BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images may include (but are not limited to):

- BLACK BORDERS
- TEXT CUT OFF AT TOP, BOTTOM OR SIDES
- FADED TEXT
- ILLEGIBLE TEXT
- SKEWED/SLANTED IMAGES
- COLORED PHOTOS
- BLACK OR VERY BLACK AND WHITE DARK PHOTOS
- GRAY SCALE DOCUMENTS

IMAGES ARE BEST AVAILABLE COPY.

As rescanning documents *will not* correct images, please do not report the images to the Image Problem Mailbox.

Exhibit A

Communication

THE JOURNAL OF BIOLOGICAL CHEMISTRY
Vol. 274, No. 7, Issue of February 12, pp. 3923–3926, 1999

The *in Vitro* Ligation of Bacterially Expressed Proteins Using an Intein from *Methanobacterium thermoautotrophicum**

(Received for publication, November 23, 1998, and in revised form, December 18, 1998)

Thomas C. Evans, Jr., Jack Benner, and Ming-Qun Xu‡

From New England Biolabs, Inc., Beverly, Massachusetts 01915-5599

The smallest known intein, found in the ribonucleoside diphosphate reductase gene of *Methanobacterium thermoautotrophicum* (*Mth* RIR1 intein), was found to splice poorly in *Escherichia coli* with the naturally occurring proline residue adjacent to the N-terminal cysteine of the intein. Splicing proficiency increased when this proline was replaced with an alanine residue. However, constructs that displayed efficient N- and C-terminal cleavage were created by replacing either the C-terminal asparagine or N-terminal cysteine of the intein, respectively, with an alanine. Furthermore, these constructs were used to specifically generate complementary reactive groups on protein sequences for use in ligation reactions. Reaction between an intein-generated C-terminal thioester on *E. coli* maltose-binding protein (43 kDa) and an intein-generated cysteine at the N terminus of either T4 DNA ligase (56 kDa) or thioredoxin (12 kDa) resulted in the ligation of the proteins through a native peptide bond. Thus the smallest of the known inteins is capable of splicing and its unique properties extend the utility of intein-mediated protein ligation to include the *in vitro* fusion of large, bacterially expressed proteins.

Inteins (1), the protein equivalent of the self-splicing RNA introns, catalyze their own excision from a precursor protein with the concomitant fusion of the flanking protein sequences, known as exteins (reviewed in Refs. 2–4). Almost 100 inteins have been identified (5)[1] and can be grouped into three classes: 1) the inteins containing a homing endonuclease between the two splicing domains, 2) the mini-inteins, which lack the homing endonuclease, and 3) a newly described trans-splicing intein (6).

Of the mini-inteins, the smallest is the 134-amino acid intein found in the ribonucleoside diphosphate reductase gene of *Methanobacterium thermoautotrophicum* (*Mth* RIR1 intein; Ref. 7). This intein may be close to the minimum amino acid sequence needed to promote splicing, and interestingly, it has a proline residue N-terminal to the first amino acid of the intein, Pro^{-1} (see Fig. 1), which was shown to inhibit splicing in an intein found in the 69-kDa vacuolar ATPase subunit of *Saccharomyces cerevisiae* (*Sce* VMA intein; Ref. 8).

Studies into the mechanism of splicing led to the development of a protein purification system that utilized thiol-induced cleavage of the peptide bond at the N terminus of the *Sce* VMA intein (9). Purification with this system generated a bacterially expressed protein with a C-terminal thioester (9). Two research groups then applied the chemistry described for native chemical ligation (10) to fuse a synthetic peptide with an N-terminal cysteine to a bacterially expressed protein possessing a C-terminal thioester (11, 12). This technique, known as intein-mediated protein ligation (IPL)[2] or also as expressed protein ligation, represented an important advance in protein semi-synthetic techniques (reviewed in Refs. 13 and 14). However, the generality of IPL was limited by the use of a synthetic peptide as a ligation partner.

We describe the next major advance in intein-mediated protein ligation, which is the modulation of the *Mth* RIR1 intein for the facile isolation of a protein with an N-terminal cysteine for use in the *in vitro* fusion of two bacterially expressed proteins. Furthermore, the *Mth* RIR1 mini-intein, the smallest known protein splicing element, was found to be capable of splicing. These results significantly expand the utility of IPL to include the labeling of extensive portions of a protein for NMR analysis and the isolation of a greater variety of cytotoxic proteins. In addition, this advance opens the possibility of labeling the central portion of a protein by ligating three fragments in succession.

EXPERIMENTAL PROCEDURES

Mth RIR1 Synthetic Gene Construction—The gene encoding the *Mth* RIR1 intein along with 5 native N- and C-extein residues (Fig. 1; Ref. 7) was constructed using 10 oligonucleotides (New England Biolabs, Beverly, MA) comprising both strands of the gene and overlapping by at least 20 base pairs. 1) 5′-TCGAGGCAACCAACCCCTGCGTATCCGGTGACACCATTGTAATGACTAGTGGCGGTCCGCGCACTGTGGCTGAACTGGAGGGCAAACCGTTCACCGCAC-3′. 2) 5′-CCGGTTGGCTGCTCGCCACAGTTGTGTACAATGAAGCCATTAGCAGTGAA TGCGCTAGCACCGTAAACAGTAGCGTCATAAACATCCTGGCGG-3′. 3) 5′-pTGATTCGCGGCTCTGGCTACCCATGCCCCTCAGGTTTCTTCCGCACCTGTGAACGTGACGTATATGATCTGCGTACACGT GAGGGTCATTGCTTACGTTT-3′. 4) 5′-pGACCCATGATCACCGTGTTCTGGTGATGGATGGTGGCCTGGAATGGCGTGCCGCGGGTGAACTGGAACGCGGCGACCGCCTGGTGATGGATGATGCAGCT-3′. 5) 5′-pGGCGAGTTTCCGGCACTGGCAACCTTCCGTGGCCTGCGTGGCGCTGGCCGCCAGGATGTTTATGACGCTACTGTTTACGGTGCTAGC-3′. 6) 5′-pGCATTCACTGCTAATGGCTTCATTGTACACAACTGTGGCGAGCAGCCAA-3′. 7) 5′-pCCAGCGCCACGCAGGCCACGGAAGGTTGCCAGTGCCGGAAACTCGCCAGCTGCATCATCCATCACCAGGCGGTCGCCGCGTTCCAGTTCACCCGCGGCAC-3′. 8) 5′-pGCCATTCCAGGCCACCATCCATCACCAGAACACGGTGATCATGGGTCAAACGTAAGCAATGACCCTCACGTGTACGCAGATCATATACGT-3′. 9) 5′-pCACGTTCACAGGTGCGGAAGAAACCTGAGGGGCATGGGTAGCCAGAGCCGCGAATCAGTGCGGTGAACGGTTTGCCCTCCAGTTCAGCCACAGTGCG-3′. 10) 5′-pCGGACCGCCACTAGTCATTACAATGGTGTCACCGGATACGCAGGGGTTGGTTGCC-3′. To ensure maximal *Escherichia coli* expression, the coding region of the synthetic *Mth* RIR1 intein

* The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

‡ To whom correspondence should be addressed. Tel.: 978-927-5054; Fax: 978-921-1350; E-mail: xum@neb.com.

[1] See also the InBase website: http://www.neb.com/neb/inteins/intein_intro.html.

[2] The abbreviations used are: IPL, intein-mediated protein ligation; MESNA, the sodium salt of 2-mercaptoethanesulfonic acid; MBP, maltose-binding protein; CBD, chitin-binding domain; M-R-B, a fusion protein consisting of maltose-binding protein-*Mth* RIR1 intein-chitin-binding domain; IPTG, isopropyl-β-D-thiogalactopyranoside; PAGE, poly-acrylamide gel electrophoresis.

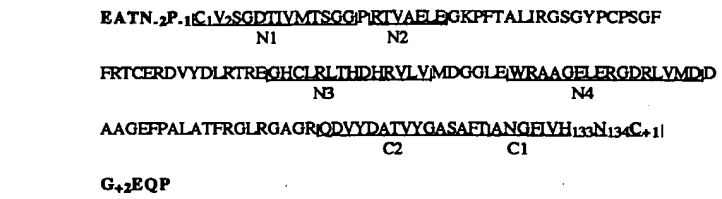


FIG. 1. ***Mth* RIR1 intein amino acid sequence.** Amino acid sequence of the *Mth* RIR1 intein with 5 native N- and C-extein residues (in *bold type*). Conserved regions of the splicing domains, N1, N2, N3, N4, C1, and C2 (22), are underlined and enclosed by *vertical bars*. The N-extein residue adjacent to the first amino acid of the intein is labeled −1 and numbering proceeds toward the N terminus of the protein (*i.e.* $N_{-2}P_{-1}$-*intein*). The intein residues are numbered sequentially starting with the N-terminal amino acid (C_{+1}). C-extein amino acids are numbered beginning with the residue immediately following the intein (*i.e.* *intein*-$C_{+1}G_{+2}$).

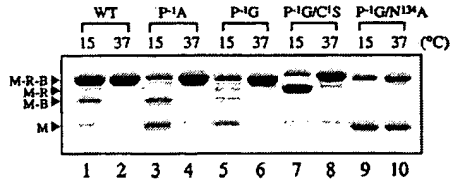


FIG. 2. **Splicing and cleavage activity of the *Mth* RIR1 intein.** Mutants of the *Mth* RIR1 intein with 5 native N- and C-terminal extein residues were induced at either 15 or 37 °C. The intein was expressed as a fusion protein (M-R-B, 63 kDa) consisting of N-terminal maltose-binding protein (*M*, 43 kDa), the *Mth* RIR1 intein (*R*, 15 kDa), and at its C terminus was the chitin-binding domain (*B*, 5 kDa). *Lanes 1* and *2*, M-R-B with the unmodified *Mth* RIR1 intein. Note the small amount of spliced product (*M-B*, 48 kDa). *Lanes 3* and *4*, *Mth* intein with Pro^{-1} replaced with Ala. Both spliced product (*M-B*) and N-terminal cleavage product (*M*) are visible. *Lanes 5* and *6*, replacement of Pro^{-1} with Gly showed some splicing as well as N- and C-terminal cleavage (*M* and *M-R*, respectively). *Lanes 7* and *8*, the Pro^{-1} to Gly and Cys^{1} to Ser double mutant ($P^{-1}G/C^{1}S$) displayed induction temperature-dependent C-terminal cleavage (*M-R*) activity. *Lanes 9* and *10*, the Pro^{-1} to Gly and Asn^{134} to Ala double mutant ($P^{-1}G/N^{134}A$) possessed only N-terminal cleavage activity producing M. The *Mth* intein or *Mth* intein-CBD fusion is not visible in this figure.

incorporates 61 silent base mutations in 48 of the 134 codons. The oligonucleotides were annealed by mixing at equimolar ratios (400 nM) in a ligation buffer (50 mM Tris-HCl, pH 7.5, containing 10 mM $MgCl_2$, 10 mM dithiothreitol, 1 mM ATP, and 25 μg of bovine serum albumin) followed by heating to 95 °C. After cooling to room temperature, the annealed and ligated oligonucleotides were inserted into the *Xho*I and *Age*I sites of pMYB5 (New England Biolabs), replacing the *Sce* VMA intein and creating the plasmid pMRB8P.

Mutagenesis of the Mth RIR1 Intein—The unique *Xho*I and *Spe*I sites flanking the N-terminal splice junction and the unique *BsrG*I and *Age*I sites flanking the C-terminal splice junction allowed substitution of amino acid residues by linker replacement. Pro^{-1}, the proline residue preceding the intein in pMRB8P, was substituted with alanine or glycine to yield pMRB8A and pMRB8G1, respectively. Substitution of Pro^{-1}-Cys^{1} with Gly-Ser or Gly-Ala yielded pMRB9GS and pMRB9GA, respectively. Replacing Asn^{134} with Ala in pMRB8G1 resulted in pMRB10G. The following linkers were used for substitution of the native amino acids at the splice junctions. Each linker was formed by annealing two synthetic oligonucleotides as described above. Pro^{-1}-Ala linker: 5′-TCGAGGCAACCAACGCATGCGTATCCGGTGACACCATTGTAATGA-3′ and 5′-CTAGTCATTACAATGGTGTCACCGGATACGCATGCGTTGGTTGCC-3′. Pro^{-1}-Gly linker: 5′-TCGAGGGCTGCGTATCCGGTGACACCATTGTAATGA-3′ and 5′-CTAGTCATTACAATGGTGTCACCGGATACGCAGCCC-3′. Pro^{-1} → Gly/Cys^{1} → Ser linker: 5′-TCGAGGGCATCGAGGCAACCAACGGATCCGTATCCGGTGA CACCATTGTAATGA-3′ and 5′-CTAGTCATTACAATGGTGTCACCGGATACGGATCCGTTGGTTGCCTCGATGCCC-3′. Pro^{-1} → Gly/Cys^{1} → Ala linker: 5′-TCGAGGGCATCGAGGCAACCAACGGCGCCGTATCCGGTGACACCATTGTAATGA-3′ and 5′-CTAGTCATTACAATGGTGTCACCGGATACGGCGCCGTTGGTTGCCTCGATGCCC-3′. Asn^{134} → Ala linker: 5′-GTACACGCATGCGGCGAGCAGCCCGGGA-3′ and 5′-CCGGTCCCGGGCTGCTCGCCGCATGCGT-3′. pBRL-A was constructed by substituting the MBP and the CBD coding regions in pMRB9GA with the CBD and the T4 DNA ligase coding regions, respectively, subcloned from the pBYT4 plasmid.[3]

Protein Splicing Studies—ER2566 cells (11) containing the appropriate plasmid were grown in LB broth containing 100 μg/ml ampicillin at 37 °C to an A_{600} of 0.5–0.8. Protein synthesis was induced by addition of 0.5 mM IPTG and proceeded at 15 °C overnight or at 37 °C for 2 h. Cell extracts were visualized on 12% Tris-glycine gels (Novex Experimental Technology, San Diego, CA) stained with Coomassie Brilliant Blue.

Protein Purification with the N-terminal Cleavage Construct—Purification was as described previously for the *Sce* VMA and *Mxe* GyrA inteins (9, 11). Briefly, ER2566 cells (11) containing the appropriate plasmid were grown at 37 °C in LB broth containing 100 μg/ml ampicillin to an A_{600} of 0.5–0.6 followed by induction with IPTG (0.5 mM). Induction was either overnight at 15 °C or for 3 h at 30 °C. The cells were pelleted by centrifugation at 3,000 × *g* for 30 min followed by resuspension in buffer A (20 mM Tris-HCl, pH 7.5, containing 500 mM NaCl). The cell contents were released by sonication. Cell debris was removed by centrifugation at 23,000 × *g* for 30 min, and the supernatant was applied to a column packed with chitin resin (bed volume, 10 ml) equilibrated in buffer A. Unbound protein was washed from the column with 10 column volumes of buffer A. Thiol reagent-induced cleavage was initiated by rapidly equilibrating the chitin resin in buffer B (20 mM Tris-HCl, pH 8, containing 500 mM NaCl and 100 mM 2-mercaptoethanesulfonic acid (MESNA)). The cleavage reaction proceeded overnight at 4 °C, after which the protein was eluted from the column.

Protein Purification with the C-terminal Cleavage Construct—Protein purification was performed as described above with buffer A replaced by buffer C (20 mM Tris-HCl, pH 8.5, containing 500 mM NaCl) and buffer B replaced by buffer D (20 mM Tris-HCl, pH 7.0, containing 500 mM NaCl). Also, following equilibration of the column in buffer D the cleavage reaction proceeded overnight at room temperature. Protein concentrations were determined using the Bio-Rad protein assay.

Protein-Protein Ligation Using IPL—Freshly isolated thioester-tagged protein was mixed with freshly isolated protein containing an N-terminal cysteine residue (starting concentration, 1–200 μM). The solution was concentrated with a Centriprep 3 or Centriprep 30 apparatus (Millipore Corporation, Bedford, MA) then with a Centricon 3 or Centricon 10 apparatus to a final concentration of 0.15–1.2 mM for each protein. Ligation reactions proceeded overnight at 4 °C and were visualized using SDS-PAGE with 12% Tris-glycine gels (Novex Experimental Technology, San Diego, CA) stained with Coomassie Brilliant Blue.

Factor Xa Cleavage of MBP-T4 Ligase Fusion Protein and Protein Sequencing—2 mg of ligation reaction involving MBP and T4 DNA ligase was bound to 3 ml of amylose resin (New England Biolabs) equilibrated in buffer A (see above). Unreacted T4 DNA ligase was rinsed from the column with 10 column volumes of buffer A. Unligated MBP and the MBP-T4 DNA ligase fusion protein were eluted from the amylose resin using buffer E (20 mM Tris-HCl, pH 7.5, containing 500 mM NaCl and 10 mM maltose). Overnight incubation of the eluted protein with a 200:1 protein:bovine factor Xa (New England Biolabs) ratio (w/w) at 4 °C resulted in the proteolysis of the fusion protein and regeneration of a band on SDS-PAGE gels that ran at a molecular weight similar to T4 DNA ligase. N-terminal amino acid sequencing of the proteolyzed fusion protein was performed on a Procise 494 protein sequencer (PE Applied Biosystems, Foster City, CA).

RESULTS

Splicing and Cleavage Activity of the Mth RIR1 Intein—The splicing activity of the *Mth* RIR1 intein with its 5 native N- and C-extein residues was investigated by expressing it as an in-frame fusion between *E. coli* maltose-binding protein (15) and the chitin-binding domain (16) from *Bacillus circulans*. In this protein context splicing products were detected (Fig. 2, *lane 1*), although the majority of the protein remained in the precursor form (M-R-B). Splicing proficiency was increased by mutating the Pro^{-1} to an Ala (Fig. 2, *lane 3*). Furthermore, the Pro^{-1} → Ala or Pro^{-1} → Gly mutants also displayed cleavage at the N- and C-terminal junctions of the intein (Fig. 2, *lanes 3* and *5*). The identity of splicing and cleavage products were confirmed by Western blot analysis using anti-MBP and anti-CBD polyclonal antibodies (data not shown).

The cleavage and/or splicing activity of the M-R-B precursor

[3] R. Chong, unpublished data.

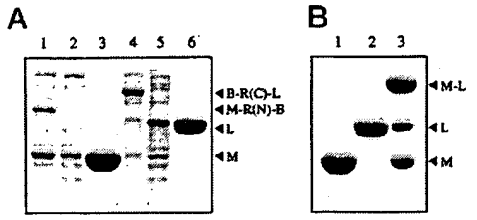


FIG. 3. **Protein purification and ligation.** *A*, thiol-inducible *Mth* intein construct (*R(N)*) for purification of MBP (*M*, 43 kDa) with a C-terminal thioester. *Lane 1*, ER2566 cells transformed with pMRB10G following IPTG induction. *Lane 2*, cell extract after passage over a chitin resin. Note that *M-R(N)-B* binds to the resin. *Lane 3*, fraction 3 of the elution from the chitin resin following overnight incubation at 4 °C in the presence of 100 mM MESNA. T4 DNA ligase (*L*, 56 kDa) purification using the C-terminal cleavage *Mth* intein construct (*R(C)*). *Lane 4*, IPTG induced ER2566 cells containing pBRL-A. *Lane 5*, cell extract after application to a chitin resin. *B-R(C)-L* binds to the resin. *Lane 6*, elution of T4 DNA ligase with an N-terminal cysteine after overnight incubation at room temperature in pH 7 buffer. *B*, ligation of MBP to T4 DNA ligase. *Lane 1*, thioester-tagged MBP. *Lane 2*, T4 DNA ligase with an N-terminal cysteine. *Lane 3*, ligation reaction of MBP (0.8 mM) with T4 DNA ligase (0.8 mM), generating *M-L*, after overnight incubation at 4 °C.

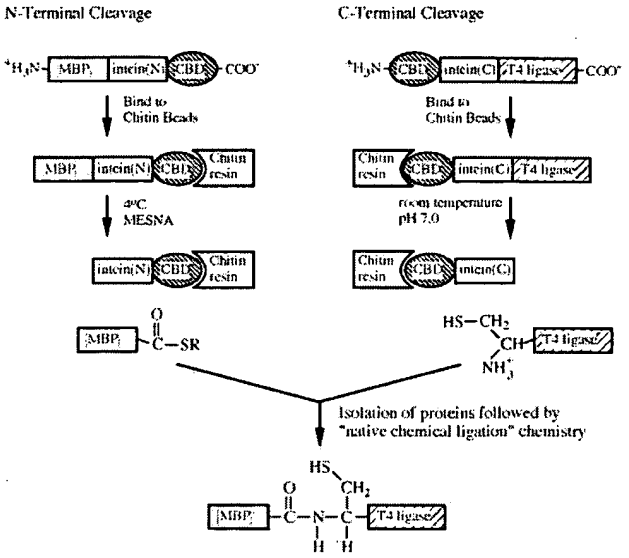


FIG. 4. **IPL pathway.** The modified *Mth* RIR1 intein was used to purify both MBP with a C-terminal thioester and T4 DNA ligase with an N-terminal cysteine. The *Mth* intein for N-terminal cleavage, intein(N), carried the $Pro^{-1} \rightarrow Gly/Asn^{134} \rightarrow Ala$ double mutation. The full-length fusion protein consisting of MBP-intein(N)-CBD was separated from cell extract by binding the CBD portion of the protein to a chitin resin. Overnight incubation in the presence of 100 mM MESNA induced cleavage of the peptide bond prior to the N terminus of the intein and created a thioester on the C terminus of MBP. The C-terminal cleavage vector, intein(C), had the $Pro^{-1} \rightarrow Gly/Cys^{1} \rightarrow Ala$ double mutation. The precursor CBD-intein(C)-T4 DNA ligase was isolated from induced *E. coli* cell extract by binding to a chitin resin as described for N-terminal cleavage. Fission of the peptide bond following the C-terminal residue of the intein resulted in the production of T4 DNA ligase with an N-terminal cysteine. Ligation occurred when the proteins containing the complementary reactive groups were mixed and concentrated, resulting in a native peptide bond between the two reacting species.

was more proficient when protein synthesis was induced at 15 °C than when the induction temperature was raised to 37 °C (Fig. 2). Replacement of Pro^{-1} with a Gly and Cys^{1} with a Ser resulted in a double mutant, M-R-B ($Pro^{-1} \rightarrow Gly/Cys^{1} \rightarrow Ser$), which showed only *in vivo* C-terminal cleavage activity when protein synthesis was induced at 15 °C but not at 37 °C (Fig. 2, *lanes 7* and *8*). Another double mutant, M-R-B ($Pro^{-1} \rightarrow Gly/Cys^{1} \rightarrow Ala$) displayed slow cleavage, even at 15 °C, which allowed the accumulation of substantial amounts of the precursor protein (data not shown) and showed potential for use as a C-terminal cleavage construct for protein purification.

Purification Using C- and N-terminal Cleavage Activity—The C- and N-terminal cleavage constructs of the *Mth* RIR1 intein were used to purify T4 DNA ligase or thioredoxin with an N-terminal cysteine or MBP with a C-terminal thioester. Two C-terminal cleavage constructs, pBRL-A and pBRT (Fig. 3, data not shown for pBRT), resulted in the isolation of 4–6 mg/liter cell culture and 5–10 mg/liter cell culture of T4 DNA ligase and thioredoxin, respectively. These proteins possessed N-terminal cysteine residues based on amino acid sequencing following the ligation reaction (see below under "Intein-mediated Protein Ligation").

Conversely, an intein with only N-terminal cleavage activity was generated by changing Pro^{-1} to Gly and the C-terminal Asn^{134} to an Ala creating M-R-B ($Pro^{-1} \rightarrow Gly$, $Cys^{1} \rightarrow Ser$). N-terminal cleavage products were detected when protein synthesis was induced at both 15 and 37 °C (Fig. 2, *lanes 9* and *10*). However, more precursor accumulated at the higher induction temperature. The remaining precursor protein could undergo thiol-mediated cleavage with reagents such as dithiothreitol or MESNA and could be used to purify thioester-tagged proteins as described previously (Fig. 3 and Refs. 11 and 12).

Intein-mediated Protein Ligation—IPL reactions consisted of mixing freshly purified MBP with T4 DNA ligase or thioredoxin (Fig. 4 and "Experimental Procedures"). Ligation was monitored by the appearance of an extra band on SDS-PAGE (Fig. 3 and data not shown for thioredoxin) corresponding to the predicted molecular weight of the ligation product. Typical ligation efficiencies ranged from 20–60%.

A factor Xa site in MBP that exists 5 amino acids N-terminal from the site of fusion (17) allowed amino acid sequencing through the ligation junction (see "Experimental Procedures"). The sequence obtained was NH_2-TLEGCGEQPTGXLK-COOH, which matched the last 4 residues of MBP (TLEG) followed by a linker sequence (CGEQPTG) and the start of T4 DNA ligase (ILK). During amino acid sequencing, the cycle expected to yield an isoleucine did not have a strong enough signal to assign it to a specific residue, so it was represented as an *X*. The cysteine was identified as the acrylamide alkylation product.

DISCUSSION

The C-terminal cleavage activity of the mutated *Mth* RIR1 intein advanced IPL technology by providing a means to isolate proteins possessing an N-terminal cysteine to act as substrates in the *in vitro* fusion of large, bacterially expressed proteins. Initially, an intein that cleaves *in vivo* was tested for the ability to generate a protein with an N-terminal cysteine. However, the side chain of the N-terminal cysteine residue appeared to be modified *in vivo* by an unidentified pathway (data not shown). Although this problem could be circumvented using a protease to cut on the N-terminal side of a cysteine residue, concern over nonspecific proteolysis and the need to remove the protease after cleavage limited its usefulness. Interestingly, C-terminal cleavage using the *Mth* RIR1 intein appeared to protect the cysteine residue until it could be released *in vitro*. A recently developed *Sce* VMA intein with thiol-inducible C-terminal cleavage activity could not be used because it would undergo splicing instead of cleavage with an N-terminal cysteine on the target protein (18).

The concentration dependence of the ligation reaction was probably due to the need to increase the ligation reaction rate to effectively compete with thioester hydrolysis, which would prevent ligation. Protein fusion occurred at 20–40% efficiency at 6.5–8.5 mg/ml of each reactant (data not shown), although

greater extents of reaction (50–60%, Fig. 3) were observed at higher protein concentrations. Many proteins can exist in solution at the lower concentrations, indicating that IPL will be useful for a wide range of applications. However, these conditions are problematic for some proteins, and future work may determine procedures that will lower this concentration requirement.

N-terminal amino acid sequencing through the ligation junction demonstrated that the two proteins were fused tail-to-head in a continuous polypeptide chain and had not fused to form an unusual branched structure. Furthermore, these data reinforce past studies reporting that a native peptide bond is formed using native chemical ligation chemistry (10) because the polypeptide sequencing reaction requires a peptide bond between amino acid residues.

Previously, studies with the *Sce* VMA intein reported that splicing was inhibited when a proline replaced the naturally occurring glycine at the −1 position (8). However, the *Mth* RIR1 intein has a naturally occurring proline at this position and was thought to be able to splice with this unique amino acid. The low splicing activity of the *Mth* RIR1 intein shows that it is capable of splicing but that it may not be folding properly when expressed in *E. coli*. Alternatively, this intein may require more native extein sequence than provided or require a cofactor such as a prolyl isomerase to promote proficient splicing activity.

The *Mth* RIR1 intein primary sequence was compared with the amino acid sequence and crystal structure of another mini-intein, the *Mxe* GyrA intein (19, 20). Most of the amino acids that form two α-helices and a disordered region in the *Mxe* GyrA intein appeared to be missing in the *Mth* RIR1 intein. The α-helical and disordered regions were previously found not to be required for splicing of the *Ssp* DnaB intein (21), and this portion of the protein may only serve as a linker. The small size of this region in the *Mth* RIR1 intein may decrease its stability and may account for some of its induction temperature-dependent activity.

The mechanism of the induction temperature-dependent splicing and cleavage activity has yet to be determined, but it may be due to reactions occurring at the C terminus of the intein. C-terminal cleavage was more severely affected by induction temperature than N-terminal cleavage activity (Fig. 2). It is also possible that the *Mth* RIR1 intein could be misfolding in *E. coli* when induced at the higher temperature, an interesting possibility considering that *M. thermoautotrophicum* is a thermophilic bacteria.

In conclusion, this report demonstrated that the smallest known intein, the *Mth* RIR1 intein, along with its 5 native extein residues was capable of splicing. Furthermore, this intein was capable of generating both thioester-tagged proteins and proteins with an N-terminal cysteine. The latter was of particular importance because it facilitated the next major advance in intein-mediated protein ligation, which is the fusion of two large, bacterially expressed proteins. This paves the way for greater freedom in the labeling of proteins for NMR analysis, the isolation of cytotoxic proteins, and in the future the controlled fusion of three bacterially expressed proteins.

Acknowledgments—We thank Bill Jack, Inca Ghosh, Francine Perler, Eric Adam, Lixin Chen, Maurice Southworth, Shoarong Chong, Eric Cantor, Chudi Guan, Richard Whitaker, Marilena Hall, and Fana Mersha for valuable discussions and assistance; Shaorong Chong for the gift of the pBYT4 plasmid; Eric Adam and Sanjay Kumar for assistance in amino acid alignments of the *Mth* RIR1 intein; and Don Comb for support and encouragement.

REFERENCES

1. Perler, F. B., Davis, E. O., Dean, G. E., Gimble, F. S., Jack, W. E., Neff, N., Noren, C. J., Thorner, J., and Belfort, M. (1994) *Nucleic Acids Res.* **22,** 1125–1127
2. Perler, F. B., Xu, M.-Q., and Paulus, H. (1997) *Curr. Opin. Chem. Biol.* **1,** 292–299
3. Perler, F. B. (1998) *Cell* **92,** 1–4
4. Xu, M.-Q., and Perler, F. B. (1996) *EMBO J.* **15,** 5146–5153
5. Perler, F. B. (1999) *Nucleic Acids Res.* **27,** in press
6. Wu, H., Hu, Z., and Liu, X. Q. (1998) *Proc. Natl. Acad. Sci. U. S. A.* **95,** 9226–9231
7. Smith, D. R., Doucette-Stamm, L. A., Deloughery, C., Lee, H., Dubois, J., Aldredge, T., Bashirzadeh, R., Blakely, D., Cook, R., Gilbert, K., Harrison, D., Hoang, L., Keagle, P., Lumm, W., Pothier, B., Qiu, D., Spadafora, R., Vicaire, R., Wang, Y., Wierzbowski, J., Gibson, R., Jiwani, N., Caruso, A., Bush, D., Safer, H., Patwell, D., Prabhakar, S., McDougall, S., Shimer, G., Goyal, A., Pietrokovski, S., Church, G. M., Daniels, C. J., Mao, J.-I., Rice, P., Nolling, J., and Reeve, J. N. (1997) *J. Bacteriol.* **179,** 7135–7155
8. Chong, S., Williams, K. S., Wotkowicz, C., and Xu, M.-Q. (1998) *J. Biol. Chem.* **273,** 10567–10577
9. Chong, S., Mersha, F. B., Comb, D. G., Scott, M. E., Landry, D., Vence, L. M., Perler, F. B., Benner, J., Kucera, R. B., Hirvonen, C. A., Pelletier, J. J., Paulus, H., and Xu, M. Q. (1997) *Gene (Amst.)* **192,** 271–281
10. Dawson, P. E., Muir, T. W., Clark-Lewis, I., and Kent, S. B. (1994) *Science* **266,** 776–779
11. Evans, J., T. C., Benner, J., and Xu, M.-Q. (1998) *Protein Sci.* **7,** 2256–2264
12. Muir, T. W., Sondhi, D., and Cole, P. A. (1998) *Proc. Natl. Acad. Sci. U. S. A.* **95,** 6705–6710
13. Gimble, F. S. (1998) *Chem. Biol.* **5,** R251-R256
14. Holford, M., and Muir, T. W. (1998) *Structure* **6,** 945–949
15. Duplay, P., Bedouelle, H., Fowler, A., Zabin, I., Saurin, W., and Hofnung, M. (1984) *J. Biol. Chem.* **259,** 10606–10613
16. Watanabe, T., Ito, Y., Yamada, T., Hashimoto, M., Sekine, S., and Tanaka, H. (1994) *J. Bacteriol.* **176,** 4465–4472
17. Maina, C. V., Riggs, P. D., Grandea, A. G., III, Slatko, B. E., Moran, L. S., Tagliamonte, J. A., McReynolds, L. A., and Guan, C. D. (1988) *Gene (Amst.)* **74,** 365–373
18. Chong, S., Montello, G. E., Zhang, A., Cantor, E. J., Liao, W., Xu, M. Q., and Benner, J. (1998) *Nucleic Acids Res.* **26,** 5109–5115
19. Klabunde, T., Sharma, S., Telenti, A., Jacobs, W. R. J., and Sacchettini, J. C. (1998) *Nat. Struct. Biol.* **5,** 31–36
20. Telenti, A., Southworth, M., Alcaide, F., Daugelat, S., Jacobs, J., William, R., and Perler, F. B. (1997) *J. Bacteriol.* **179,** 6378–6382
21. Wu, H., Xu, M.-Q., and Liu, X.-Q. (1998) *Biochim. Biophys. Acta* **1387,** 422–432
22. Pietrokovski, S. (1998) *Protein Sci.* **7,** 64–71

EXHIBIT B

ELSEVIER

Biochimica et Biophysica Acta 1387 (1998) 422–432

BIOCHIMICA ET BIOPHYSICA ACTA BBA

Protein *trans*-splicing and functional mini-inteins of a cyanobacterial dnaB intein

Hong Wu [a], Ming-Qun Xu [b], Xiang-Qin Liu [a,*]

[a] *Biochemistry Department, Dalhousie University, Halifax, Nova Scotia B3H 4H7, Canada*
[b] *New England Biolabs, Inc., 32 Tozer Road, Beverly, MA 01915, USA*

Received 9 March 1998; accepted 8 May 1998

Abstract

A 429 aa theoretical intein is encoded in the *dnaB* gene (DNA helicase) of the cyanobacterium *Synechocystis* sp. strain PCC6803. This intein is shown to be capable of protein splicing with or without its native exteins when tested in *E. coli* cells. A centrally located 275 amino acid sequence (residues 107–381) of this intein can be deleted without loss of the protein splicing activity, resulting in a functional mini-intein of 154 aa in size. Efficient in vivo protein *trans*-splicing was observed when this mini-intein was split into a 106 aa N-terminal fragment containing intein motifs A and B, and a 48 aa C-terminal fragment containing intein motifs F and G. These results indicate that the N- and C-terminal regions of the *Ssp* DnaB intein, whether covalently linked with each other or not, can come together through non-covalent interaction to form a protein splicing domain that is functionally sufficient and structurally independent from the centrally located endonuclease domain of the intein.

Keywords: Intein; Protein *trans*-splicing; Splicing domain; DNA helicase; (Cyanobacterium)

1. Introduction

An intein is a protein sequence embedded in-frame within a precursor protein sequence and excised during a maturation process termed protein splicing [1,2]. Protein splicing is a post-translational event involving precise excision of the intein sequence and concomitant ligation of the flanking sequences (N- and C-exteins) by a normal peptide bond [3–5]. Approximately 50 intein-coding sequences have been found in over 20 different genes distributed among the nuclear and organellar genomes of eukaryotes, archaebacteria (archaea), and eubacteria, suggesting a wide distribution of inteins (see the Intein Registry at http://www.neb.com/neb/inteins.html). Known inteins share little overall sequence identity, except between closely related inteins found at the same insertion site in homologous proteins of different organisms [6]. Nevertheless, a number of short sequence motifs (sequence blocks A to H) have been recognized that show a low but significant degree of conservation among inteins [6,7], suggesting similar structure, function, and evolutionary origin of different inteins. Molecular mechanisms of protein splicing have been studied, and they involve N→S (or N→O) acyl shift at the splice sites [5,8,9], formation of a branched intermediate [10,11], and cyclization of an invariant Asn residue at the C-terminus of intein to form succinimide [12], leading to excision of the intein and ligation of the exteins.

* Corresponding author. Fax: +1-902-494-1355; E-mail: pxqliu@is.dal.ca

0167-4838/98/$ – see front matter
PII: S0167-4838(98)00157-5

Many inteins are bi-functional elements, possessing protein splicing activity as well as endonuclease activity involved in intein homing (mobility) [13–15]. Structure determination of the *Sce* VMA1 intein by X-ray crystallography has revealed a two-domain structure, with domain I consisting the N- and C-terminal regions of the intein sequence and domain II formed by the middle part of the intein sequence [16]. Domain II was suggested to be the endonuclease domain, with domain I (or a part of it) corresponding to the splicing domain. Such a bipartite structure may be applicable to inteins in general, as has been suggested by other studies including mutagenesis studies [17,18] and sequence statistical modeling [19,20]. Functional studies of some inteins have confirmed such a two-domain model. Deleting the endonuclease domain of the *Sce* VMA1 intein and the *Mtu* RecA intein have produced mini-inteins that are capable of protein splicing [21,22]. The *Mxe* GyrA intein naturally lacks an endonuclease domain and was shown to be capable of protein splicing [23]. These findings suggest that the N- and C-terminal regions of an intein make up a functional splicing domain and that the centrally located endonuclease domain is not required for splicing. But differences seem to exist among different inteins. In the *Psp* Pol-1 intein, for example, deletions (gaps) of different sizes in the endonuclease domain all led to inactivation of the splicing activity [24].

These observations raise questions of whether the above findings are applicable to other inteins, whether the endonuclease domain (if present) and the native exteins of a particular intein may play a role in the correct folding and function of the splicing domain, and whether the N- and C-terminal sequences of an intein can come together and assemble properly in the absence of both the endonuclease domain and a covalent linkage between them. We have investigated the *Ssp* DnaB intein to address some of these questions. The *Ssp* DnaB intein is a 429 amino acid (aa) intervening sequence encoded in the *dnaB* (DNA helicase) gene of the cyanobacterium *Synechocystis* sp. strain PCC6803, and it has been recognized as a theoretical intein based on the presence of intein-like sequence motifs [25]. In addition to residues and motifs associated with a protein splicing domain, this intein has sequence motifs for an endonuclease domain. The *Ssp* DnaB intein is also related to a homologous intein in *Rhodothermus marinus* likely through recent intein homing [29]. Here we demonstrate that this intein is capable of protein splicing with or without its native exteins when tested in *Escherichia coli* cells. A centrally located 275 aa sequence of this intein, corresponding to the entire endonuclease domain, could be deleted without losing the protein splicing activity. The resulting mini-intein was split into two fragments, and efficient protein *trans*-splicing was observed. These results indicate that the N- and C-terminal regions of the *Ssp* DnaB intein, whether physically linked or not, can come together to form a protein splicing domain that is functionally sufficient and structurally independent from the centrally located endonuclease domain.

2. Materials and methods

2.1. DNA cloning

The complete *dnaB* coding sequence (2616 base pair long) was isolated from total DNA of *Synechocystis* sp. strain PCC6803. This was done by specifically amplifying the *dnaB* DNA in a polymerase chain reaction (PCR) using the thermostable DNA polymerase Pfu (Stratagene) and a pair of oligonucleotide primers: 5′-CGGAATTCCATATGGCTG-CTAACCCTGCCCT-3′ and 5′-CGCTGCAGGAT-CCTAGTAATCATTACTTCGTTGC-3′. Plasmid pTS1 was constructed by inserting a 1796 base pair (bp) *Nco*I–*Bam*HI DNA fragment (blunt ended) of the *dnaB* gene into the expression plastid vector pET-32 (Novagen) at its *Bam*HI site (blunt ended), so that the *dnaB* coding sequence was in-frame with the upstream vector-encoded sequence of thioredoxin, polyhistidine tag, and the S tag (a peptide sequence: KE-TAAAKFERQHMDS). Plasmid pTS2 was constructed by in-frame deletion of a 174 bp fragment from the 3′ end of the *dnaB* coding sequence. pTS3 was constructed by inserting the 1796 bp *Nco*I–*Bam*HI DNA fragment of the *dnaB* coding sequence into the expression plasmid vector pET-16b (Novagen) at its *Nco*I–*Bam*HI site.

Deletion plasmids (pTS1-1 through pTS1-5) were all derived from pTS1. A nested deletion method was used to construct pTS1-1, pTS1-2, pTS1-4 and pTS1-5. In this method, the pTS1 DNA was first cleaved at

a *Spe*I site located near the middle of the DnaB intein coding sequence, the resulting linear DNA was subjected to progressive deletion from both ends using exonucleases provided in the Nested Deletion kit from Pharmacia, and these were followed by ligation of the two ends to re-circularize the DNA. Plasmid pTS1-3 was derived from pTS1 by a PCR-mediated deletion method. First, a linear DNA fragment was amplified from the circular pTS1 DNA in a polymerase chain reaction (PCR), using a mixture of the thermostable *Taq* DNA polymerase (Promega) and Vent DNA polymerase (New England Biolabs), and from a pair of oligonucleotide primers: 5′-GGATCCCAATTGTCACCAGAAATAGAAAAG-3′, and 5′-ACTCCCCAATTGTAAAGAGGAGCTTTC-3′. The amplified linear DNA fragment was then circularized to form pTS1-3.

Plasmid pMST was derived from a previously constructed pMYT1 plasmid that encodes a tripartite fusion protein consisting of *E. coli* Maltose-binding protein, Yeast *Sce* VMA1 intein, and *E. coli* Thioredoxin [9]. The yeast intein coding sequence (Y) was replaced with the coding sequence of the *Ssp* DnaB mini-intein from pTS1-3 to produce pMST. Plasmid pMST-n was derived from pMST by introducing a translation termination codon into the *Ssp* DnaB mini-intein coding sequence. Plasmid pMST-split was derived from pMST by introducing a cassette of (termination codon)–(Shine–Dalgarno sequence)–(initiation codon) into the mini-intein coding sequence. This was achieved by using a PCR-mediated method. First, a linear DNA fragment was amplified from the circular pMST DNA in a polymerase chain reaction, using the Advantage cDNA polymerase mix (Clontech) and a pair of oligonucleotide primers: 5′-GGAGGTTTAAAATATGTCACCAGAAATAGAAAAGTTGTC-3′, and 5′-CCTCATTATAATTGTAAAGAGGAGCTTTCTA-3′. The amplified linear DNA molecule was then circularized to form pMST-split.

2.2. Protein production and splicing in E. coli cells

E. coli cells transformed with individual recombinant plasmids of interest were grown in liquid Luria Broth medium at 37°C to late log phase (A_{600}, 0.5). IPTG was added to a final concentration of 0.8 mM to induce production of the recombinant proteins, and the induction was continued for 3 h at 37°C or 25°C as specified. Cells were lysed in SDS-containing gel loading buffer in a boiling water bath before SDS–polyacrylamide gel electrophoresis. In isolating proteins containing poly-histidine tag, cells were lysed in a denaturing buffer (50 mM NaH_2PO_4, 10 mM Tris–HCl, 8 M urea, pH 8.0), and the target proteins were selectively precipitated by using the TALON metal affinity resin (Clontech) which binds the poly-histidine peptide sequence. In detecting proteins containing specific sequences, Western blottings were carried out by using an S protein (Novagen) that specifically recognizes the S-tag sequence, an anti-MBP antiserum (New England Biolabs) that specifically recognizes the maltose binding protein sequence, an anti-Trx antiserum (American Diagnostica) that specifically recognizes thioredoxin, and an anti-intein antiserum that was raised against the *Ssp* DnaB intein sequence. Estimations of the amount of protein in individual protein bands were carried out by using a gel documentation system (Gel Doc 1000 coupled with Molecular Analyst software, Bio-Rad).

3. Results

3.1. Protein splicing of the complete Ssp DnaB intein

The *Ssp* DnaB intein was tested for protein splicing activity in *E. coli* cells. The complete *Ssp dnaB* gene was isolated from total DNA of *Synechocystis* sp. strain PCC6803 by selectively amplifying the *dnaB* gene in a polymerase chain reaction (PCR). We were unable to clone the entire *Ssp dnaB* gene in an expression plasmid vector, presumably due to toxicity of the gene product (a DNA helicase) in the *E. coli* cell. Clones containing partial *Ssp dnaB* gene were readily obtained, and they included three recombinant plasmids (pTS1, pTS2 and pTS3), each encoding a fusion protein consisting of the complete intein sequence flanked by various amount of extein sequences and tag sequences (Fig. 1A). Production of each fusion protein is controlled by an IPTG-inducible T7 promoter.

Each recombinant plasmid was introduced into *E. coli* cells to produce the corresponding fusion protein and to observe possible protein splicing products (Fig. 1B). In cells containing plasmid pTS1, three

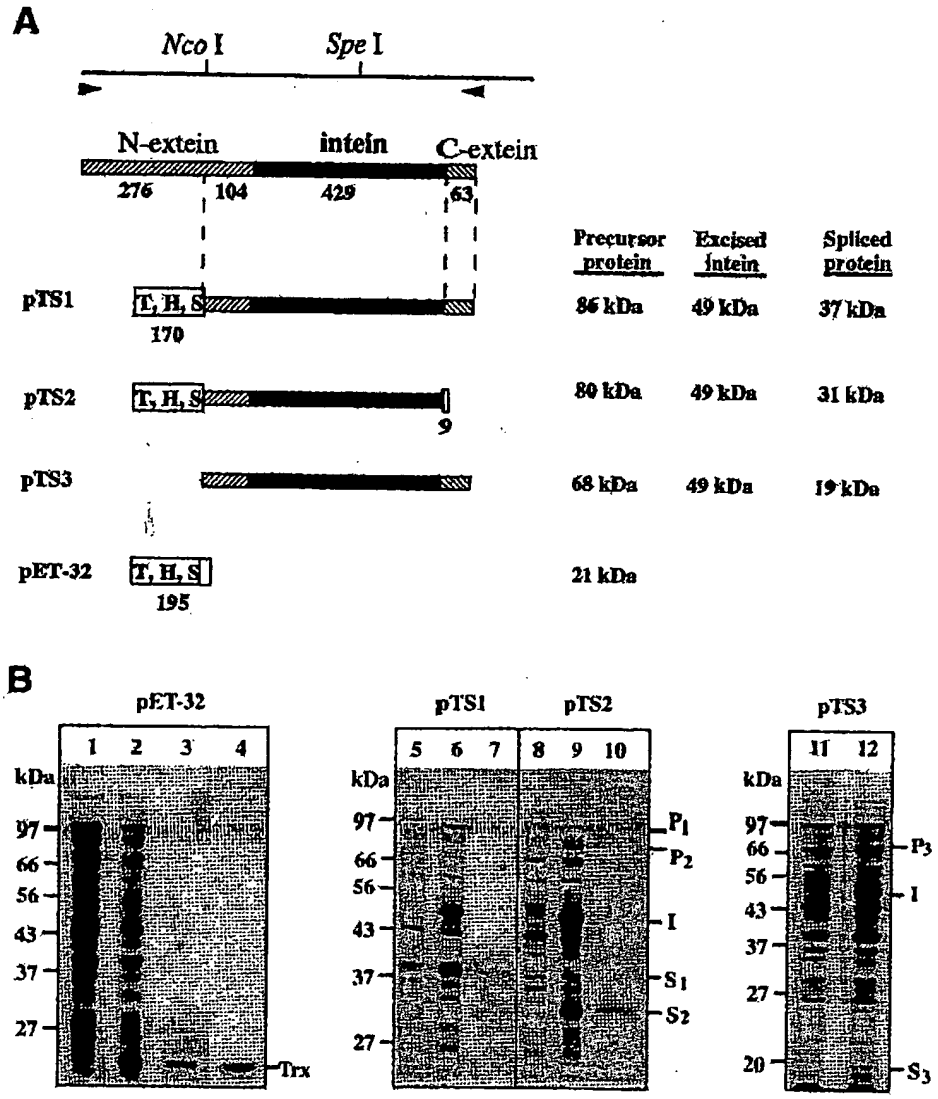
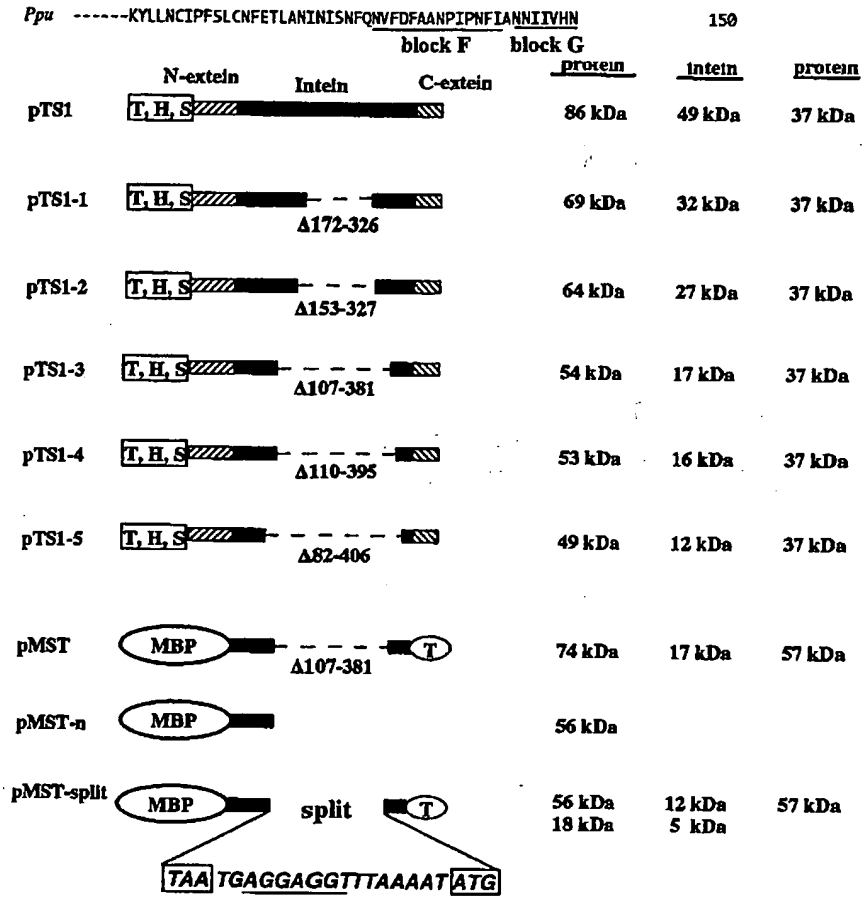


Fig. 1. Protein splicing with complete *Ssp* DnaB intein. (A) Schematic illustration of fusion protein construct. The top line shows restriction sites and oligonucleotide primers (arrowheads) used in this study. The DnaB intein (solid box) and extein (hatched box) sequences are fused with vector-encoded sequences (open boxes), with the number of residues shown for each sequence. T, H, and S stand for thioredoxin, poly-histidine tag, and S tag, respectively. For each construct (pTS1, pTS2, pTS3, and pET-32 as a control), calculated molecular masses are listed for the predicted precursor protein, the excised intein and the spliced protein. (B) Observation of protein splicing. *E. coli* cells containing the specified plasmid were induced at 25°C, and the induced proteins were analyzed by SDS–polyacrylamide gel electrophoresis and Coomassie blue staining. Lanes 1, 5, 8, and 11: before induction. Lanes 2, 6, 9, and 12: after induction. Lanes 3, 7, and 10: proteins isolated by using metal affinity resin that recognizes the poly-histidine tag. Lane 4: Western blot using the S protein that recognizes the S tag. Letters P, I, S, and Trx mark positions of precursor protein, excised intein, spliced protein, and thioredoxin, respectively.

protein products were observed. Their sizes corresponded closely to the predicted sizes of a precursor protein (86 kDa), a spliced protein (37 kDa), and an excised intein (49 kDa), respectively. Three protein products were also observed in cells containing plasmid pTS2, and their sizes corresponded well with the predicted sizes of a precursor protein (80 kDa), a spliced protein (31 kDa), and an excised intein (49 kDa), respectively. Similarly, cells containing plasmid pTS3 produced three proteins corresponding to a precursor (68 kDa), a spliced protein (19 kDa), and an excised intein (49 kDa), respectively. In addition



B

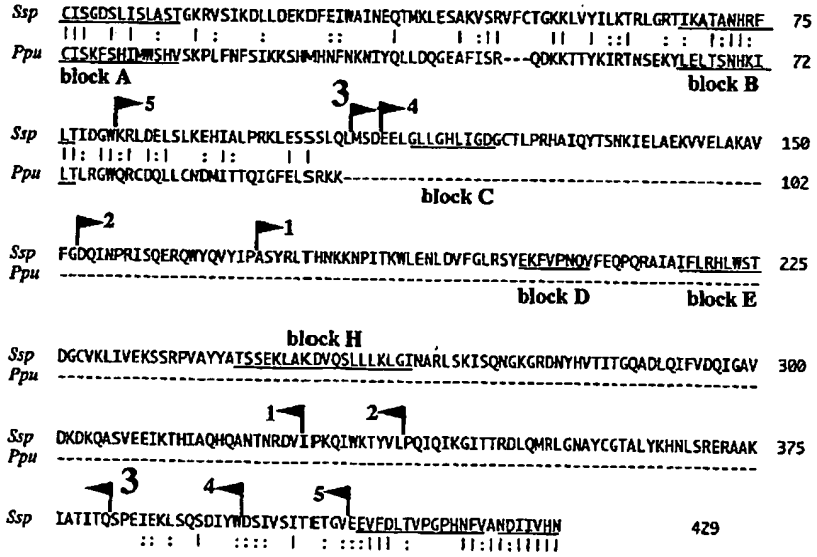


Fig. 2. Construction of mini-inteins and a split intein. (A) Schematic illustration of fusion proteins encoded in the corresponding plasmids. pTS1 encodes the complete intein sequence, while pTS1-1 through pTS1-5 encode intein sequences with deletions of various sizes. Deleted areas of the intein are marked by dashed lines, and their boundaries are specified by the numbers. In pMST-split, the DNA sequence of a small insertion is shown, with the termination codon *TAA* and the initiation codon *ATG* enclosed in boxes and the Shine–Dalgarno sequence underlined. In each case, the DnaB intein (solid box) and extein (hatched box, if present) sequences are fused in-frame with vector-encoded sequences (open boxes or circles). T, H, S, and MBP stand for thioredoxin, poly-histidine tag, S tag, and maltose binding protein, respectively. Calculated molecular masses of the predicted protein products are listed. (B) Intein sequence of deletion constructs. The *Ssp* DnaB intein sequence (*Ssp*) is aligned with the *Ppu* DnaB intein sequence (*Ppu*). Blocks A to H are conserved intein motifs [6]. Symbols: – represents gaps introduced to optimize the alignment; | and : mark positions of identical and similar amino acids, respectively. Flags 1 through 5 mark deletion boundaries of the deletion construct pTS1-1 through pTS1-5, respectively. For example, the upstream and downstream deletion boundaries of pTS1-1 are marked by the right-pointing flag 1 and the left-pointing flag 1, respectively.
←

to identification by size, the precursor and spliced protein bands were further identified by selective binding to metal affinity resin (property of poly-histidine tag) and to the S protein (property of the S tag). The intein band was identified by its size, by the fact that its size was not affected by changing the extein sequences, and also by Western blot using an intein-specific antiserum (described later in Fig. 3A).

3.2. *Protein splicing of Ssp DnaB intein containing deletions*

A series of deletion mutations were introduced in the *Ssp* DnaB intein coding sequence (Fig. 2A), using either a nested deletion method or a PCR-mediated method. The intein sequence and deletion boundaries of these deletion mutations are shown in Fig. 2B. As a guide in constructing the deletion mutations, the *Ssp* DnaB intein sequence was aligned with the related but smaller intein sequence of *Porphyra purpurea* chloroplast (*Ppu* DnaB intein). Previously recognized putative intein motifs (sequence blocks A to H) were also taken into consideration. In particular, one deletion mutation (pTS1-3) was constructed to have its deleted area matching closely to the sequence gap between the *Ssp* DnaB intein and the *Ppu* DnaB intein. Deletion constructs pTS1-1 through pTS1-5 were made in the expression plasmid vector pET32 in the same configuration as the control plasmid pTS1 (see Fig. 1A).

These recombinant plasmids were introduced into *E. coli* cells to produce corresponding fusion proteins and to observe protein splicing products (Fig. 3A). Presence of protein splicing is indicated by the production of a spliced protein and an excised intein in addition to a precursor protein. Identification of each precursor protein, excised intein and spliced protein was based on a combination of two observations: (1) the protein's apparent size, which should match closely to its predicted size; and (2) the predicted presence or absence of specific sequences or sequence tags, which were confirmed either by Western blots using antisera against specific sequences, by binding to the S protein in a Western blot (a property of the S-tag sequence), or by binding to a metal affinity resin (a property of their poly-histidine sequence).

It is apparent from Fig. 3A that protein splicing occurred in cells containing pTS1-1, pTS1-2, and pTS1-3, as indicated by the production of the spliced protein (ligated exteins) in each case. Protein splicing appeared less efficient with the deletion constructs pTS1-1, pTS1-2, and pTS1-3, when compared to the control plasmid pTS1 containing the complete intein. Western blot (Fig. 3A, lanes 21–24) was used to estimate the amount of spliced protein as a percentage of the total (spliced protein plus precursor protein). This percentage of spliced protein was estimated to be 78%, 15%, 23%, and 41% for pTS1, pTS1-1, pTS1-2, and pTS1-3, respectively. In constructs pTS1-4 and pTS1-5 that contain larger deletions in the intein sequence, a spliced protein was not observed, indicating the absence of a detectable amount of protein splicing. In cells containing pTS1-3, the excised intein band was observed on the stained gel and readily identified on Western blot using an anti-intein antiserum. In cells containing pTS1-1 or pTS1-2, an excised intein is also expected because of the observed spliced protein, but the excised intein was not apparent from the stained gel, indicating a low level of accumulation. A minor band was observed just beneath the precursor pro-

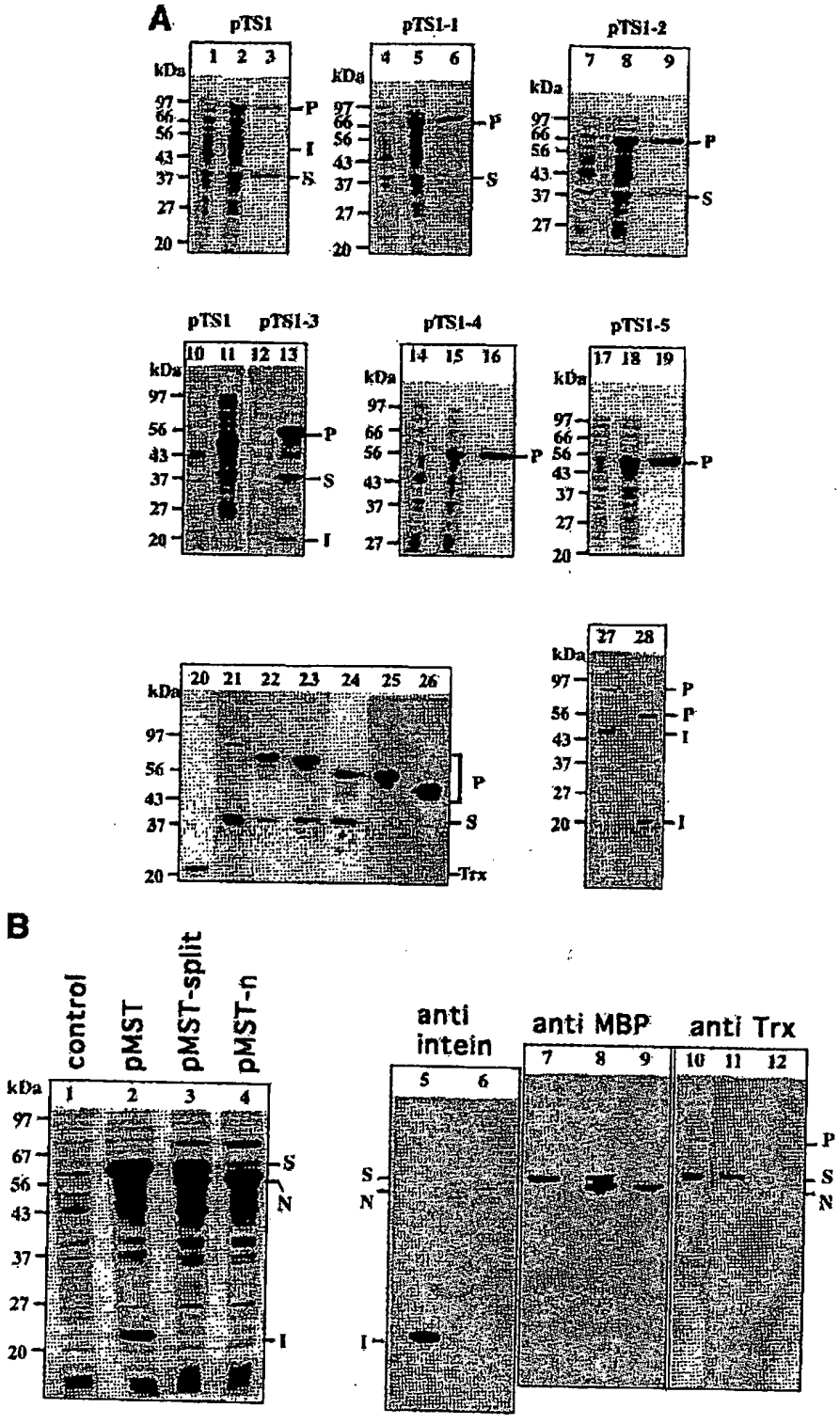

A
pTS1
pTS1-1
pTS1-2
pTS1-3
pTS1-4
pTS1-5
kDa
P
S
I
Trx
B
control
pMST
pMST-split
pMST-n
anti intein
anti MBP
anti Trx
N

Fig. 3. Protein splicing of mini-inteins and a split intein. *E. coli* cells containing individual plasmids were induced at 25°C (37°C for pMST), and the induced proteins were analyzed by SDS–polyacrylamide gel electrophoresis followed by Coomassie blue staining or Western blotting. (A) Protein splicing of mini-inteins. Lanes 1, 4, 7, 10, 12, 14, and 17: before induction. Lanes 2, 5, 8, 11, 13, 15, and 18: after induction. Lanes 3, 6, 9, 16, and 19: proteins isolated by using metal affinity resin that recognizes the poly-histidine tag. Lanes 20–26: Western blot using the S protein that recognizes the S tag, on total proteins of cells containing plasmid pET-32 (lane 20), pTSI (lane 21), pTS1-1 (lane 22), pTS1-2 (lane 23), pTS1-3 (lane 24), pTS1-4 (lane 25), or pTS1-5 (lane 26). Lanes 27 and 28: Western blot using an intein-specific antiserum, on total proteins of cells containing pTS1 (lane 27) or pTS1-3 (lane 28). Letters P, I, S, and Trx mark positions of precursor protein, excised intein, spliced protein, and thioredoxin, respectively. (B) Protein splicing of a split intein. Lanes 1–4: proteins of control cell (before induction, lane 1) and cells containing pMST (lane 2), pMST-split (lane 3), and pMST-n (lane 4). Lanes 5 and 6: Western blot using anti-intein antiserum on proteins of lanes 2 and 3, respectively. Lanes 7, 8, and 9: Western blot using anti-MBP antiserum on proteins of lanes 2, 3, and 4, respectively. Lanes 10, 11, and 12: Western blot using anti-thioredoxin antiserum on proteins of lanes 2, 3, and 4, respectively. Letters P, S, and I mark positions of precursor protein, spliced protein, and excised intein, respectively. Letter N marks protein product of pMST-n.

tein band for each of the deletion constructs on the Western blot (Fig. 3A, lanes 22–26), suggesting a cleavage or breakdown product.

3.3. Protein trans-splicing of a split mini-intein

Plasmid pMST was constructed at first, which permitted better identification of the protein products. The intein sequence in pMST has exactly the same deletion as in pTS1-3, but the native extein sequences (except five residues proximal to the intein) are replaced by the *E. coli* maltose binding protein at the N-terminus and *E. coli* thioredoxin at the C-terminus (Fig. 2A). Retaining the five proximal native extein residues on both sides of the intein is to avoid potential disturbance of the intein active site by proximal foreign extein residues. Cells containing pMST showed efficient protein splicing, with both the spliced protein and the excised intein readily observed and identified (Fig. 3B). In addition to the predicted size, the spliced protein was recognized by antiserum against MBP, by antiserum against thioredoxin, but not by antiserum against the intein, all as expected. The excised intein was recognized only by an antiserum against the intein. There was very little accumulation of the precursor protein, indicating that the protein splicing is more efficient in the pMST construct than it is in the pTS1-3 construct (comparing lane 2 of Fig. 3B with lane 13 of Fig. 3A), with the two constructs having identical intein sequences but different flanking (extein) sequences. Also, pMST showed efficient protein splicing both at 25°C and at 37°C, while pTS1-3 showed little or no protein splicing at 37°C (data not shown).

In testing for protein *trans*-splicing, plasmid pMST-split was constructed from pMST by splitting the functional mini-intein in pMST into two parts (Fig. 2A). This was achieved by inserting in the intein coding sequence a cassette consisting of (translation termination codon)–(Shine–Dalgarno sequence)–(translation initiation codon). The resulting pMST-split is essentially a two-gene operon, with the first gene (gene I) encoding the N-extein sequence plus the N-terminal sequence of the intein, and with the second gene (gene II) encoding the C-terminal sequence of the intein plus the C-extein sequence. A control plasmid pMST-n was also constructed by inserting only a translation termination codon in the intein coding sequence, without introducing the Shine–Dalgarno sequence and the translation initiation codon.

In *E. coli* cells containing the pMST-split plasmid, production of a spliced protein was observed (Fig. 3B, lane 3). This spliced protein is, by design, identical to the spliced protein produced from pMST through protein *cis*-splicing. In addition to the expected size, the spliced protein from pMST-split was recognized by antiserum against MBP, by antiserum against thioredoxin, but not by antiserum against the intein, all as expected. In addition to the spliced protein, the protein product of the first gene (gene I) of the two-gene operon was also accumulated. This protein (labeled N in Fig. 3B) was first identified by its size, which is the same as the protein product of pMST-n. Also as expected, this protein was recognized by antiserum against MBP, by antiserum against the intein, but not by antiserum against thioredoxin. A protein product of the second gene (gene II) of the two-gene operon was not observed. The accumulation of gene I protein but not

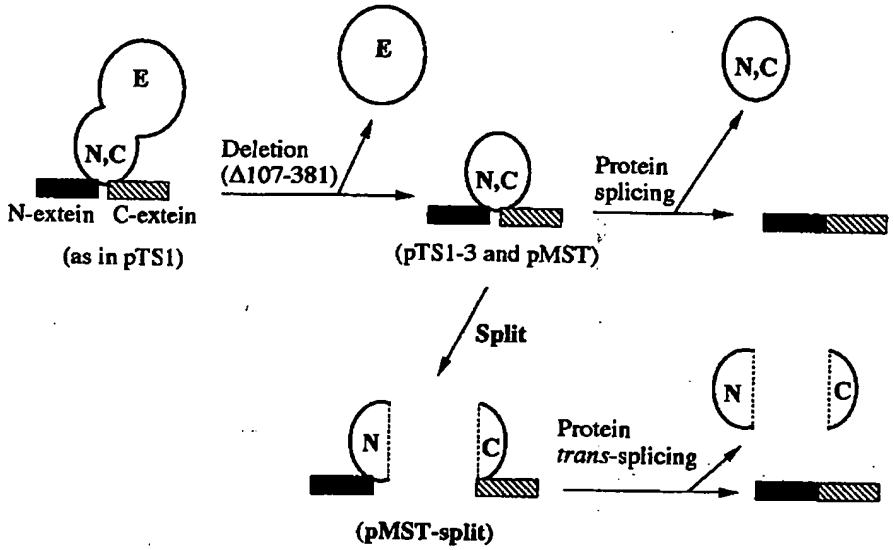


Fig. 4. Functional mini-intein and split intein. A model of the *Ssp* DnaB intein is shown to relate its function with its structure. Structural domain [N,C] contains the N-terminal region (N, residues 1–107, approximately) and the C-terminal region (C, residues 381–429, approximately) of the intein sequence, while domain E represents the endonuclease domain in the middle part of the intein sequence. Domain [N,C] is a functional splicing domain that is sufficient for protein splicing, as was demonstrated by the deletion constructs pTS1-3 and pMST. This splicing domain can function without its N part and its C part being covalently linked in a single polypeptide, as was demonstrated by the split intein construct pMST-split that undergoes protein *trans*-splicing.

gene II protein indicates that gene I protein was produced in molar excess (relative to gene II protein), probably due to a less than 100% translational coupling between gene I and gene II. Two excised intein fragments, predicted to be 12 and 5 kDa, respectively, were not observed, most likely due to their small sizes, weak recognition by the anti-intein antiserum that was raised against a continuous intein, and/or rapid degradation in the *E. coli* cell.

4. Discussion

As illustrated in Fig. 4, we have shown that the endonuclease domain of the *Ssp* DnaB intein sequence is not required for protein splicing and that the two terminal regions of the intein need not be covalently linked for protein splicing to occur. In the mini-intein constructs pTS1-3 and pMST, the centrally located 275 aa sequence was deleted, producing a functional mini-intein of just 154 aa in size. In the split intein construct pMST-split, the N- and C-terminal sequences of the mini-intein could be produced as two separate pieces without losing the splicing function. Although a crystal structure is not available for the *Ssp* DnaB intein, a splicing domain and an endonuclease domain could be inferred from the above findings and from known crystal structures of the *Sce* VMA1 intein and the *Mxe* GyrA intein. Statistical modeling has produced sequence alignments among the *Ssp* DnaB intein, the *Sce* VMA1 intein, and the *Mxe* GyrA intein along with many other inteins [19,20], suggesting a structural resemblance among different inteins. Based on these sequence alignments, the functional mini-intein (154 aa) derived from the *Ssp* DnaB intein corresponds to a major part (approximately 70%) of domain I (splicing domain) of the *Sce* VMA1 intein, while the 275 aa sequence that was deleted from the *Ssp* DnaB intein corresponds to the entire domain II (endonuclease domain) plus a part of domain I of the *Sce* VMA1 intein [16]. The crystal structure of *Mxe* GyrA intein showed a β-core formed by the N- and C-terminal sequences of the intein, with the middle part of the intein sequence forming a disordered region and two α helices that extend from the β-core [26]. Based on this crystal structure and a sequence alignment between *Ssp* DnaB intein and *Mxe* GyrA intein [19], the functional mini-intein (154 aa) derived from the *Ssp* DnaB

intein corresponds to the entire β-core of the *Mxe* GyrA intein, while lacking most of the disordered region and the α helices present in the *Mxe* GyrA intein. In the split intein (pMST-split) derived from the *Ssp* DnaB intein, the N-terminal intein fragment corresponds to the N-terminal nine β strands (β1 through β9) of the *Mxe* GyrA intein, while the C-terminal fragment corresponds to the C-terminal three β strands (β10, β11, β12) of the *Mxe* GyrA intein.

Efficient protein *trans*-splicing of the split mini-intein construct pMST-split indicates that the N-terminal fragment (106 aa) and the C-terminal fragment (48 aa) of the *Ssp* DnaB intein can come together to form a functional splicing domain without assistance of either the endonuclease domain or a covalent linkage between them. Crystal structures of both the *Sce* VMA1 intein and the *Mxe* GyrA intein revealed non-covalent interactions between the N- and C-terminal sequences of the intein [16,26]. In the *Mxe* GyrA intein structure, for example, a region (β10) of the C-terminal sequence meets with several regions (β4, β5, β6) of the N-terminal sequence to form antiparallel β-strands and three-stranded mixed β-sheets [26]. Our observation of *trans*-splicing with the split *Ssp* DnaB mini-intein suggests that non-covalent interactions between the N- and C-terminal sequences of this intein are sufficient to bring the two sequences into correct assembly or folding for the *trans*-splicing to occur. The N-extein (maltose binding protein) and the C-extein (thioredoxin) are, by design, two separate and stable structural domains. They are not known to interact with each other and therefore are unlikely to contribute to the reassembly of the two intein fragments. Consistent with this, protein *trans*-splicing was also observed after replacing these non-native exteins with native exteins of *Ssp* DnaB intein (data not shown).

The demonstration of in vivo protein *trans*-splicing may also have implications on intein evolution. In addition to losing its endonuclease domain in evolution, an intein may further lose its continuity by a split in its intein-coding sequence and still retain the ability to produce a mature (functional) host protein through protein *trans*-splicing. In agreement with this, the *Ssp* DnaE intein has recently been found as a naturally occurring split intein that does protein *trans*-splicing [30]. In a study of the *Psp* Pol-1 intein, two intein fragments (precursors) were produced in separate cells and subsequently reconstituted in vitro to initiate *trans*-splicing, with the objective of controlling protein splicing by intein fragment reassembly [24]. Unlike the in vivo *trans*-splicing of *Ssp* DnaB intein, the in vitro *trans*-splicing of *Psp* Pol-1 intein required a denaturation–renaturation step in the presence of urea. This difference mostly likely reflects the different conditions employed in the two studies, because in vivo reassembly of the two intein fragments may occur before the co-expressed fragments misfold and may also be assisted by the protein folding machinery of the cell. There is another perhaps more fundamental difference between the *Ssp* DnaB intein and the *Psp* Pol-1 intein. In the *Psp* Pol-1 intein, all tested combinations of intein fragments that resulted in a deletion (gap) in the endonuclease domain failed to show protein *trans*-splicing. This is in contrast to the *Ssp* DnaB intein in which the entire endonuclease domain is not required for the *trans*-splicing. This difference also exist in protein *cis*-splicing, because the *Psp* Pol-1 intein, unlike the *Ssp* DnaB intein, failed to show *cis*-splicing when deletions were made in the endonuclease domain. The *Mtu* RecA intein was shown recently to support protein *trans*-splicing in *E. coli* cells [27], and in vitro reconstitution of the engineered intein fragments also resulted in *trans*-splicing after renaturation from 6 M urea [28].

Efficient protein splicing of the *Ssp* DnaB mini-intein (construct pMST) is consistent with a two-domain structure of this intein. It also indicates that the splicing domain is structurally and functionally independent from both the endonuclease domain and the native exteins. Similar observations have been made with other inteins, but there appear to be differences among different inteins including the *Ssp* DnaB intein. In the *Mtu* RecA intein, the entire endonuclease domain could be deleted while retaining lower levels of splicing activity. In the *Sce* VMA1 intein, a portion of the endonuclease domain could be replaced with a linker polypeptide without abolishing the splicing function, but deleting the entire endonuclease domain led to a loss of protein splicing [21]. In the *Psp* Pol-1 intein, deletions of different sizes in the endonuclease domain all led to inactivation of the splicing activity. Among the naturally occurring mini-inteins lacking an endonuclease domain, only

the *Mxe* GyrA intein has been shown to splice, and a linker insertion (in place of the endonuclease domain) as well as the native N-extein are required for splicing [23]. These differences among inteins suggest that the endonuclease domain (if present) and the native exteins of some (but not all) inteins may play a role in the correct folding and function of the splicing domain. In this respect, it is noted that the *Ssp* DnaB mini-inteins (pTS1-1 and pTS1-2) that have partial endonuclease domain sequences showed less efficient protein splicing in comparison to the mini-intein pTS1-3 that lacks the entire endonuclease domain. The *Ssp* DnaB mini-intein flanked by native extein sequences (in pTS1-3) also showed less efficient protein splicing when compared to an identical mini-intein flanked by non-native exteins (in pMST). These observations suggests that the partial endonuclease domain and native extein sequences may have actually interfered with the proper folding of the precursor protein or the splicing domain. The *Ssp* DnaB mini-intein is remarkably efficient in *cis*- and *trans*-splicing, in that the splicing reactions are not accompanied by either upstream or downstream cleavage, and that the splicing reactions are not dependent on either a spacer sequence or native exteins. In contrast, many other inteins either requires a spacer sequence [21], is dependent on a native extein [23], or undergoes significant amount of cleavages (e.g., Refs. [9,10,22,27]). This suggests that the *Ssp* DnaB intein is intrinsically more efficient than the other inteins in heterologous system, perhaps owing to a more robust structure or folding ability. This difference among inteins is both interesting and of practical importance in engineering inteins for various applications.

Acknowledgements

We thank Dr. Donald Comb for his generous support and Zhuma Hu for her technical assistance. This work was supported by a grant from the Medical Research Council of Canada.

References

[1] F.B. Perler, E.O. Davis, G.E. Dean, F.S. Gimble, W.E. Jack, N. Neff, C.J. Noren, J. Thorner, M. Belfort, Nucleic Acids Res. 22 (1994) 1125–1127.
[2] F.B. Perler, Cell 92 (1998) 1–4.
[3] M.J. Colston, E.O. Davis, Mol. Microbiol. 12 (1994) 359–363.
[4] A.A. Cooper, T.H. Stevens, Trends Biochem. Sci. 20 (1995) 351–356.
[5] M.-Q. Xu, F.B. Perler, EMBO J. 15 (1996) 5146–5153.
[6] F.B. Perler, G.J. Olsen, E. Adam, Nucleic Acids Res. 25 (1997) 1087–1093.
[7] S. Pietrokovski, Protein Sci. 3 (1994) 2340–2350.
[8] Y. Shao, M.Q. Xu, H. Paulus, Biochemistry 35 (1996) 3810–3815.
[9] S. Chong, Y. Shao, H. Paulus, J. Benner, F.B. Perler, J. Biol. Chem. 271 (1996) 22159–22168.
[10] M.Q. Xu, M.W. Southworth, F.B. Mersha, L.J. Hornstra, F.B. Perler, Cell 75 (1993) 1371–1377.
[11] M.Q. Xu, D.G. Comb, H. Paulus, C.J. Noren, Y. Shao, F.B. Perler, EMBO J. 13 (1994) 5517–5522.
[12] Y. Shao, M.Q. Xu, H. Paulus, Biochemistry 34 (1995) 10844–10850.
[13] F.S. Gimble, J. Thorner, Nature 357 (1992) 301–306.
[14] R.F. Doolittle, Proc. Natl. Acad. Sci. U.S.A. 90 (1993) 5379–5381.
[15] M. Belfort, R. Roberts, Nucleic Acids Res. 25 (1997) 3379–3388.
[16] X. Duan, F.S. Gimble, F.A. Quiocho, Cell 89 (1997) 555–564.
[17] M. Kawasaki, S. Nogami, Y. Satow, Y. Ohya, Y. Anraku, J. Biol. Chem. 272 (1997) 15668–15674.
[18] S. Nogami, Y. Satow, Y. Ohya, Y. Anraku, Genetics 147 (1997) 73–85.
[19] J.Z. Dalgaard, M.J. Moser, R. Hughey, I.S. Mian, J. Comput. Biol. 4 (1997) 193–214.
[20] J.Z. Dalgaard, A.J. Klar, M.J. Moser, W.R. Holley, A. Chatterjee, I.S. Mian, Nucleic Acids Res. 25 (1997) 4626–4638.
[21] S. Chong, M.-Q. Xu, J. Biol. Chem. 272 (1997) 15587–15590.
[22] V. Derbyshire, D.W. Wood, W. Wu, J.T. Dansereau, J.Z. Dalgaard, M. Belfort, Proc. Natl. Acad. Sci. U.S.A. 94 (1997) 11466–11471.
[23] A. Telenti, M. Southworth, F. Alcaide, S. Daugelat, W.R. Jacobs Jr., F.B. Perler, J. Bacteriol. 179 (1997) 6378–6382.
[24] M.W. Southworth, E. Adam, D. Panne, R. Byer, R. Kautz, F.B. Perler, EMBO J. 17 (1998) 918–926.
[25] S. Pietrokovski, Trends Genet. 12 (1996) 287–288.
[26] T. Klabunde, S. Sharma, A. Telenti, W.R. Jacobs Jr., J.C. Sacchettini, Nat. Struct. Biol. 5 (1998) 31–36.
[27] K. Shingledecker, S.-Q. Jiang, H. Paulus, Gene 207 (1998) 187–195.
[28] K.V. Mills, B.M. Lew, S.-Q. Jiang, H. Paulus, Proc. Natl. Acad. Sci. USA 95 (1998) 3543–3548.
[29] X.-Q. Liu, Z. Hu, Proc. Natl. Acad. Sci. USA 94 (1997) 7851–7856.
[30] H. Wu, Z. Hu, X.-Q. Liu, Proc. Natl. Acad. Sci. USA 95 (1998) 9226–9231.

Exhibit C

Communication

THE JOURNAL OF BIOLOGICAL CHEMISTRY
Vol. 273, No. 26, Issue of June 26, pp. 15887–15890, 1998

Protein Splicing *in Vitro* with a Semisynthetic Two-component Minimal Intein*

(Received for publication, April 9, 1998)

Belinda M. Lew‡§¶, Kenneth V. Mills‡§¶, and Henry Paulus‡‖**

From the ‡Boston Biomedical Research Institute, Boston, Massachusetts 02114, §Department of Chemistry and Chemical Biology, Harvard University, Cambridge, Massachusetts 02138, and ‖Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston, Massachusetts 02115

Protein splicing elements, or inteins, catalyze their own excision from flanking polypeptide sequences, or exteins, thereby leading to the formation of new proteins in which the exteins are linked directly by a peptide bond. A *trans*-splicing system, using separately purified and expressed N- and C-terminal intein fragments of about 100 amino acids each, fused to appropriate exteins, was recently derived from the *Mycobacterium tuberculosis* RecA intein (Mills, K. V., Lew, B. M., Jiang, S.-Q., and Paulus, H. (1998) *Proc. Natl. Acad. Sci. U. S. A.* 95, 3543–3548). We have replaced the C-terminal intein fragment of this system with synthetic peptides comprising 35–50 of the C-terminal residues of the RecA intein. The N-terminal intein fragment and the synthetic peptide were reconstituted by renaturation from guanidinium chloride. In the absence of added reductants, a disulfide-linked dimer of the N-terminal fragment and the peptide accumulated and could be induced to splice by reduction of its disulfide bond. The intermediate and spliced products were identified by polyacrylamide gel electrophoresis, mass spectrometry, and derivatization with thiol-reactive biotin followed by Western blotting with a streptavidin-enzyme conjugate. This is the first example of protein splicing involving a synthetic intein fragment and opens the way for studying the active site structure and function of the intein by the use of different synthetic peptides, including ones with non-natural amino acids.

Protein splicing is a mechanism for the post-translational processing of proteins that involves the self-catalyzed excision of an intervening polypeptide, the intein, and the subsequent formation of a new protein by joining the flanking sequences, the exteins, by a peptide bond. It involves the catalysis of three mechanistically unrelated reactions at a single catalytic center, which resides entirely within the intein (for reviews, see Refs. 1 and 2). The intein can thus be viewed as an exceedingly complex enzyme, and the investigation of the catalytic mechanisms involved in protein splicing is of great interest.

With the aim of obtaining an *in vitro* protein splicing system whose structure and function can be examined by biochemical and biophysical methods, we are developing a minimal protein splicing element from the 440-residue RecA intein of *Mycobacterium tuberculosis* by eliminating the portions of the intein that are not essential for protein splicing. Most inteins are interrupted by homing endonuclease domains, which account for about one-half of the intein sequence but can be deleted without eliminating protein splicing ability (3–5). In addition, we found that the RecA intein can be split into two fragments that can complement each other so as to promote *trans*-splicing (5). This made possible the development of an *in vitro trans*-splicing system composed of 105-residue N- and C-terminal fragments of the *M. tuberculosis* RecA intein (6). An *in vitro trans*-splicing system based on the Pol-1 intein of the hyperthermophilic archeon, *Pyrococcus* sp. GB-D, was recently described (7). The results described in this paper further advance our approach by replacing the natural C-terminal intein fragment with 35–50-residue synthetic polypeptides. The resulting semisynthetic protein splicing element was able to catalyze protein splicing with high efficiency. This exciting development will facilitate the study of the structure and catalytic function of the C-terminal portion of the intein by replacing specific residues with other natural amino acids or with unnatural amino acids and structural probes.

EXPERIMENTAL PROCEDURES

Plasmid Constructs and Protein Expression and Purification—The construction of plasmid pMU2 s/sD6, which encodes $MU_{N\Delta}$[1] (an in-frame fusion of MBP to the 105 N-terminal amino acids of the *M. tuberculosis* RecA intein, followed by the C-terminal sequence Arg-Gly-Glu-Phe) was described earlier (5). $MU_{N\Delta}$ was expressed in *Escherichia coli* DH5α and purified as described previously (6).

Peptide Synthesis—Peptides with a C-terminal amide group ranging from 33 to 52 residues in length (see Fig. 1) were synthesized by *N*-(9-fluorenyl)methoxycarbonyl chemistry on an Applied Biosystems model 431 peptide synthesizer, using 4-methyl benzhydrylamine resin (Novabiochem, San Diego, CA). *O*-trityl-protected Thr, double coupling of all Thr and Arg residues, an acetic anhydride blocking step at each cycle, and *N*-methylpyrrolidone supplemented with dimethyl sulfoxide to 10%. The peptides were cleaved from the resin and deprotected with 6.25% (w/v) phenol in thioanisole:1,2-ethanedithiol:water:trifluoroacetic acid (2:1:2:20). The peptides were purified by HPLC (Rainin), using a preparative C8 column (Vydac) and monitoring the fractions by MALDI-TOF MS.

MALDI-TOF MS—Protein samples were dialyzed overnight against water and mixed on a mass spectrometry plate with an equal volume of 2,5-dihydroxybenzoic acid (1 mg/ml) in water:isopropyl alcohol:formic acid (3:2:1). An external standard of 1 mg/ml bovine serum albumin was also prepared, and the $MU_{N\Delta}$ starting material was used as an internal mass standard. MS was performed on a Voyager RP Biospectrometry

* This work was supported by Grant R01 GM55875 from the National Institute of General Medical Sciences (to H. P.) and by a Howard Hughes Medical Institute predoctoral fellowship (to K. V. M.). The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

¶ B. M. L. and K. V. M. should be considered joint first authors.

** To whom reprint requests should be addressed: Boston Biomedical Research Inst., 20 Staniford St., Boston, MA 02114. Tel.: 617-912-0350; Fax: 617-912-0308; E-mail: paulus@bbri.harvard.edu.

[1] The abbreviations used are: $MU_{N\Delta}$, chimeric protein consisting of an N-terminal MBP fused to $U_{N\Delta}$; MBP, *E. coli* maltose-binding protein; $U_{N\Delta}$, the 105 N-terminal amino acids of the *M. tuberculosis* RecA intein, followed by the C-terminal Arg-Gly-Glu-Phe-COOH; BMCC, 1-biotinamido-4-[4′-(maleimidomethyl)-cyclohexanecarboxamido]butane; DTT, DL-1,4-dithiothreitol; GdmCl, guanidinium chloride; M, N-terminal extein containing a spacer and N-terminal MBP; MALDI-TOF, matrix-assisted laser desorption ionization-time of flight; PAGE, polyacrylamide gel electrophoresis; TBS, Tris-buffered saline; TCEP, tris(2-carboxyethyl)phosphine; HPLC, high pressure liquid chromatography; MS, mass spectrometry.

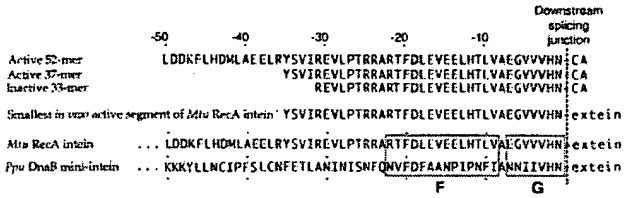


FIG. 1. **Alignment of the synthetic peptides used in this work with the C-terminal domain of the *M. tuberculosis* (*Mtu*) RecA intein, the minimal segment of the RecA intein shown to be functional *in vivo* (3), and the C-terminal domain of the *Porphyria purpurea* (*Ppu*) DnaB mini-intein (12).** Amino acid residues are numbered backwards from the downstream splicing junction, which is indicated by the *dotted line*. The *boxed regions* labeled *F* and *G* represent the conserved C-terminal intein motifs (13).

Workstation (PerSeptive Biosystems, Framingham, MA) using linear mode and a low mass gate of 2000. PerSeptive GRAMS/386 data analysis software was used, and a Savitsky-Golay 19-order smoothing function was performed on all spectra.

Reconstitution and Splicing Procedure—The standard reconstitution procedure involved mixing $MU_{N\Delta}$ and peptide at concentrations of 9 and 84 μM, respectively, unless otherwise indicated, followed by dialysis against Buffer N (20 mM sodium phosphate, pH 7.5; 6 M GdmCl; 500 mM NaCl; 1 mM EDTA) for 1 h at 4 °C, using SpectraPor 3500 MWCO dialysis tubing, followed by dialysis for 1 h against three changes of Buffer O (Buffer N without GdmCl). A sample of the dialyzed mixture was saved at 4 °C to study formation of the $MU_{N\Delta}$/peptide heterodimer, and the remainder was allowed to undergo splicing by adding TCEP to 1 mM and incubating at 25 °C for 16 h. Protein and peptide concentrations were estimated from their absorbance at 280 nm and the calculated molar absorption coefficients as described earlier (6).

Analysis of Reconstitution and Splicing Products—In some cases, samples were reduced with 1 mM TCEP and then biotinylated by treatment with 0.04 volume of 8.5 mM BMCC (Pierce) in dimethyl sulfoxide at 25 °C for 2 h. Samples were analyzed by SDS-PAGE using precast 10–20% gradient Tris/glycine gels (Owl Scientific, Cambridge, MA) and prestained protein markers (New England Biolabs, Beverly, MA), according to the method of Laemmli (8), except that DTT was omitted from the sample buffer where indicated. Gels were stained for protein with Coomassie Blue. To screen for biotinylated proteins, gels were blotted for 16 h onto nitrocellulose membranes (Schleicher and Schuell) at 36 V. The blots were soaked for 30 min in blocking buffer (1% bovine serum albumin in 20 mM Tris, pH 7.5; 150 mM NaCl), washed twice for 5 min in TBS (20 mM Tris-HCl, pH 7.5; 150 mM NaCl), and then incubated with 2 mg/ml alkaline phosphatase-conjugated streptavidin (Pierce) diluted 1:2000 in TBS for 1 h. The blots were washed twice for 5 min in TBS, and immobilized alkaline phosphatase activity was detected using 5-bromo-4-chloro-3-indolyl phosphate/nitroblue tetrazolium substrate tablets (Sigma). Gels and Western blots were scanned with a Supervista S-12 scanner (Umax Data Systems) and analyzed densitometrically using the NIH Image 1.60 program.

RESULTS

Ability of a Synthetic Peptide to Function in Protein Splicing—In the *in vitro trans*-splicing system developed earlier, 105-residue N- and C-terminal fragments of the *M. tuberculosis* RecA intein, fused to appropriate exteins, were mixed in 6 M urea or GdmCl and reconstituted by removing the denaturant by dialysis (6). The two intein fragments formed a heterodimer, which underwent efficient protein splicing under reducing conditions. Upon replacing the C-terminal intein fragment with a synthetic peptide corresponding to the 50 C-terminal intein residues, linked to Cys-Ala as the C-extein (Fig. 1), an analogous set of reactions was observed. As shown in Fig. 2 (*lane 5*), about 55% of $MU_{N\Delta}$ was converted to a 61-kDa protein, whose molecular mass was consistent with that of a disulfide-linked $MU_{N\Delta}$/peptide heterodimer. Upon addition of the reductant, TCEP, the 61-kDa protein was replaced by a new 43-kDa protein, in an amount corresponding to 50% of $MU_{N\Delta}$ and consistent in molecular mass with the putative spliced product, *i.e.* M linked to Cys-Ala (Fig. 2, *lane 4*). The overall splicing reaction proceeded somewhat more efficiently (60% yield based

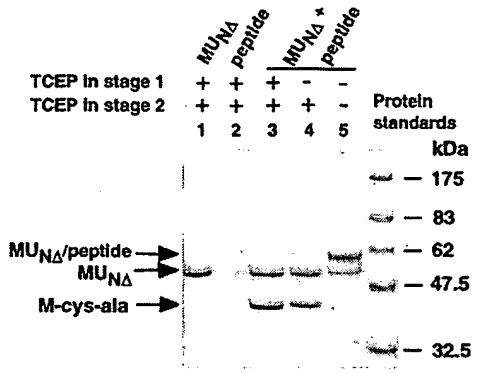


FIG. 2. **SDS-PAGE analysis of protein splicing with a synthetic peptide as the C-terminal component of the protein splicing element.** $MU_{N\Delta}$ and the 52-residue peptide were reconstituted by renaturation at 4 °C from 6 M GdmCl (stage 1), followed by incubation at 25 °C (stage 2) as described under "Experimental Procedures," except that TCEP (1 mM) was added as indicated. Control incubations were done with $MU_{N\Delta}$ (*lane 1*) and the peptide (*lane 2*) alone. DTT was omitted from the SDS-PAGE sample buffer in *lane 5*.

on $MU_{N\Delta}$) when TCEP was added at the beginning of the reconstitution procedure (Fig. 2, *lane 3*). Neither the 61-kDa intermediate nor the 43-kDa putative spliced product was observed when the peptide or $MU_{N\Delta}$ was omitted from the reaction (Fig. 2, *lanes 1* and *2*).

Identification of Intermediates and Products—The putative heterodimeric intermediate and splicing product were further characterized by MS. Samples of the starting materials, the products of renaturation in the absence of TCEP, and the splicing reaction after addition of TCEP were analyzed by MALDI-TOF MS. The major molecular species that was observed in the starting mixture corresponded to $MU_{N\Delta}$ (m/z = 55,064) and the 52-residue peptide (m/z = 6,051) (Fig. 3*B*). Upon renaturation under non-reducing conditions, an additional component was found with m/z of 61,100, in close agreement with that expected for the disulfide-linked heterodimer of $MU_{N\Delta}$ and the 52-mer (m/z = 61,113) (Fig. 3*C*). Upon addition of TCEP, the m/z = 61,100 ionic species disappeared, and three new major ionic species appeared with m/z = 43,200, 12,000, and 5,900, consistent with the predicted m/z values of 43,293, 11,964, and 5,875 for the splicing product (M-Cys-Ala), the intein fragment ($U_{N\Delta}$), and the 50-residue peptide fragment, respectively (Fig. 3*D*).

The expected product of protein splicing, M-Cys-Ala, differs by a mass of only 193 mass units, corresponding to the dipeptide Cys-Ala, from M itself, which could have been produced from cleavage at the upstream splice junction (5). We therefore used an independent chemical assay for the identification of the putative protein splicing product. Because protein splicing leads to transfer of Cys-Ala to the C terminus of M, which itself contains no Cys residues, it should be possible to distinguish between M and the splicing product, M-Cys-Ala, by a method that specifically detects thiols. The products of the splicing reaction were treated with the thiol-reactive biotin-maleimide derivative, BMCC, which should specifically label all proteins containing Cys residues, including $MU_{N\Delta}$ and the splicing product, M-Cys-Ala, but not free M. After SDS-PAGE and blotting onto nitrocellulose membranes, biotinylated proteins were detected using a streptavidin-alkaline phosphatase conjugate. In the complete splicing mixtures, a 43-kDa protein was the major biotin-labeled species (Fig. 4, *lanes 1* and *3*), whereas only $MU_{N\Delta}$ was labeled in a mixture without the 52-residue peptide (Fig. 4, *lane 5*) and neither labeled component was observed when $MU_{N\Delta}$ was omitted (Fig. 4, *lane 7*). No signal was observed with samples that had not been subjected to

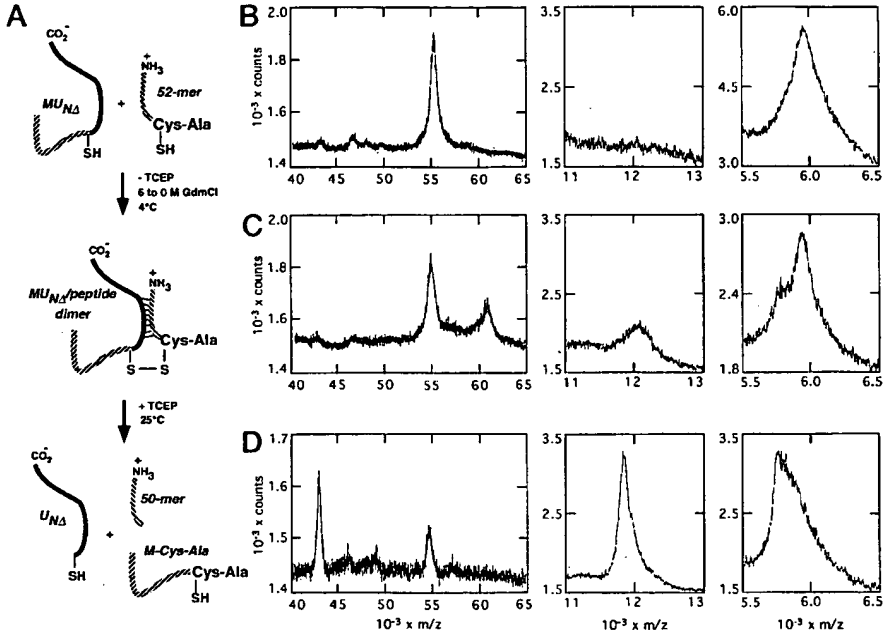


FIG. 3. **MALDI-TOF MS analysis of protein splicing involving a synthetic peptide.** Protein splicing elements were reconstituted from $MU_{N\Delta}$ and the 52-residue peptide and induced to splice as outlined on the *left* (*A*) and as described under "Experimental Procedures." Samples of the starting materials (*B*), the dialyzed reconstitution mixture (*C*), and the splicing products (*D*) were prepared for MS. *Left panels*, mass range, 40–65 kDa; *center panels*, mass range, 10–14 kDa; *right panels*, mass range, 5.5–6.5 kDa.

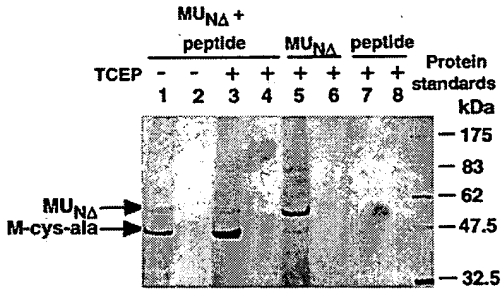


FIG. 4. **Identification of thiol-containing polypeptides by Western blotting after reaction with a biotin-maleimide derivative.** Reconstitution and splicing of $MU_{N\Delta}$ and the 52-residue peptide, biotinylation of the splicing products with BMCC, and Western blotting with a streptavidin-alkaline phosphatase conjugate were done as described under "Experimental Procedures," except that in samples 3–8, TCEP also was present during the reconstitution reaction. The samples in *lanes 2, 4, 6,* and *8* were not biotinylated.

biotinylation (Fig. 4, *lanes 2, 4, 6,* and *8*). As another control, molecular mass markers including MBP were labeled with BMCC and developed on a Western blot. All marker proteins with free thiols were labeled, whereas MBP was not (data not shown). The observation that the 43-kDa protein produced in the complete system could be biotinylated with BMCC identified it as the spliced protein, M-Cys-Ala, rather than a cleavage product such as M.

Characteristics of the Peptide-dependent Splicing Reaction—Investigation of the role of prior denaturation on the reconstitution and splicing reactions showed that significant amounts of $MU_{N\Delta}$/peptide heterodimer were formed when $MU_{N\Delta}$ and the peptide were mixed in the absence of GdmCl, together with some higher aggregates (Fig. 5, *lane 3*), but that subsequent reduction yielded little spliced product (Fig. 5, *lane 4*) compared with reaction mixtures in which $MU_{N\Delta}$ and the peptide were reconstituted under denaturing conditions (Fig. 5, *lanes 1* and *2*). It is interesting that the heterodimer formed under non-denaturing conditions (Fig. 5, *lane 3*) failed to undergo efficient splicing, suggesting that productive interaction of the intein fragments to form a functional protein splicing active center requires prior unfolding of the polypeptide chains.

The experiments described in Figs. 2–5 were carried out with a nearly 10-fold molar excess of the 52-residue peptide. When the ratio of peptide to $MU_{N\Delta}$ was varied at a constant concentration (42 μM) of $MU_{N\Delta}$, maximum conversion to spliced product (55%) was observed with an equimolar amount of peptide, suggesting a stoichiometric interaction of the two intein components (Fig. 6). The extent of conversion of $MU_{N\Delta}$ to spliced product roughly paralleled the extent of conversion to disulfide-linked heterodimer when this was measured separately (see Figs. 2, *lanes 4* and *5* and Fig. 5, *lanes 1* and *2*). The extent of conversion of $MU_{N\Delta}$ to spliced product varied from 55 to 90% (for example, compare Figs. 2 and 4).

Effect of Peptide Length on Protein Splicing—Peptides comprising fewer than the 50 C-terminal amino acids of the protein splicing element were also examined for their ability to function in protein splicing. Each peptide was present in a 3-fold molar excess with respect to $MU_{N\Delta}$. The results summarized in Fig. 7 show that a peptide corresponding to the 35 C-terminal amino acids of the intein was fully able to substitute for the 52-mer, whereas a peptide corresponding to the 31 C-terminal amino acids was inactive.

DISCUSSION

The results described in this paper demonstrate that a semisynthetic protein splicing element can effectively catalyze the complex series of reactions that lead to protein splicing. Our experimental system consisted of two fragments of an intein linked to appropriate exteins, which could be non-covalently reconstituted to form a functional protein splicing element; one intein fragment was a natural 105-residue protein segment and the other a synthetic 50-residue polypeptide. There have been other examples of active semisynthetic enzymes that can be reconstituted by the association of a natural and a synthetic fragment, the first being the reconstitution of ribonuclease S from S-peptide (residues 1–20) and S-protein (residues 21–124), which are produced by the cleavage of ribonuclease A with subtilisin (9). Replacement of the S-peptide with synthetic analogs yields functional semisynthetic ribonuclease derivatives (*e.g.* Ref. 10). Protein splicing elements should lend themselves especially well to reconstitution as semisynthetic enzymes because the protein splicing active center is composed of polypeptide sequences that correspond to the extreme ends of

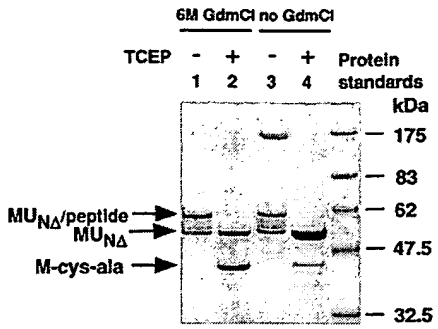


FIG. 5. **Effect of prior denaturation on reconstitution and protein splicing yield.** Reconstitution of MU$_{N\Delta}$ and the 52-residue peptide was done either in the presence of 6 M GdmCl (Buffer N) or with no GdmCl (Buffer O) as indicated and analyzed by SDS-PAGE either directly (samples 1 and 3) or after the induction of splicing with TCEP (samples 2 and 4) as described under "Experimental Procedures." DTT was omitted from the SDS-PAGE sample buffer for samples 1 and 3.

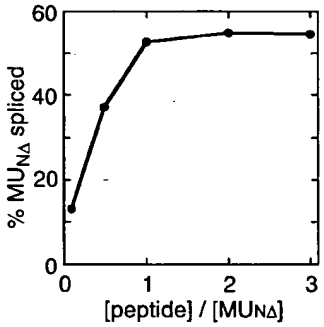


FIG. 6. **Effect of peptide concentration on splicing yield.** Various concentrations of the 52-residue peptide ranging from 4.2 to 125 μM and a constant amount of MU$_{N\Delta}$ (42 μM) were subjected to reconstitution and splicing as described under "Experimental Procedures," except that 1 mM TCEP was also present during reconstitution. The samples were subjected to SDS-PAGE, and the amount of spliced product was estimated by densitometry after staining with Coomassie Blue. The data are presented as the fraction of MU$_{N\Delta}$ converted to spliced product as a function of the molar ratio of peptide to MU$_{N\Delta}$.

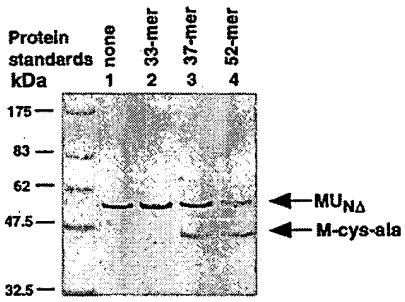


FIG. 7. **Effect of peptide length on the ability to function as a component of a protein splicing element.** MU$_{N\Delta}$ was mixed with a 3-fold molar excess of synthetic peptides corresponding to the 31, 35, or 50 C-terminal amino acids of the *M. tuberculosis* RecA intein and with C-terminal Cys-Ala as the extein, and induced to reconstitute and splice as described under "Experimental Procedures," except that 1 mM TCEP was also present during reconstitution. The samples were subjected to SDS-PAGE, followed by staining with Coomassie Blue.

the intein. Interspersed between these protein splicing sequences is often an extensive, functionally unrelated homing endonuclease domain (11), which imposes a spatial and temporal gap between their synthesis and assembly into a single functional domain. Indeed, natural N- and C-terminal fragments of the *M. tuberculosis* RecA intein, separately expressed and purified, were found to reconstitute and undergo protein splicing with high efficiency (6).

Our observation that synthetic polypeptides corresponding to between 35 and 50 of the C-terminal amino acids of the intein could effectively promote protein splicing offers an excellent opportunity for probing the structure and function of the protein splicing active center by substituting other amino acids or unnatural amino acid analogs at specific positions in the peptide. An especially attractive feature of our experimental system is that we can measure the reconstitution reaction separately from protein splicing by using mildly oxidizing conditions under which a refolded disulfide-linked heterodimer accumulates, which can subsequently be made to undergo quantitative conversion to the spliced products by reduction with TCEP. One can, therefore, study the effect of amino acid substitutions on the reconstitution reaction *per se*, *i.e.* the formation of a disulfide-linked complex, or on the protein splicing reaction *per se*, which occurs upon reduction of the disulfide-linked heterodimer. In addition, because the disulfide-linked heterodimer can be isolated as a stable protein, the structure of the protein splicing active center and its perturbation by amino acid substitution can be studied by various biophysical methods. The unusual nature of the protein splicing element as an enzyme should make such future investigations especially exciting.

One question that can be addressed immediately concerns the minimum size of the downstream intein fragment that is required for protein splicing. In a deletion analysis, Derbyshire *et al.* (3) found that protein splicing occurs *in vivo* when all but the last 35 C-terminal amino acids of the *M. tuberculosis* RecA intein are deleted but not after deletion of all but the last 31 residues. The 50-residue sequence used in most of our experiments is larger than the minimum size of the C-terminal intein fragment required for protein splicing. However, we could reconstitute a functional semisynthetic protein splicing element with a synthetic peptide corresponding to the 35 C-terminal residues of the *M. tuberculosis* RecA intein but not with one corresponding to the 31 C-terminal residues (Fig. 7). By synthesizing polypeptides of intermediate size, we should be able to define precisely the minimal length required for protein splicing.

Acknowledgments—We thank Kaori Shingledecker for help and advice, Shu-qin Jiang for preparing the plasmid, Anna Wong for synthesizing the peptides, and Paul Leavis and Betty Gowell for use of their preparative HPLC.

REFERENCES

1. Shao, Y., and Kent, S. B. H. (1997) *Chem. Biol.* **4**, 187–194
2. Perler, F., Xu, M. Q., and Paulus, H. (1997) *Curr. Opin. Chem. Biol.* **1**, 292–299
3. Derbyshire, V., Wood, D. W., Wu, W., Dansereau, J. T., Dalgaard, L. Z., and Belfort, M. (1997) *Proc. Natl. Acad. Sci. U. S. A.* **94**, 11466–11471
4. Chong, S., and Xu, M. Q. (1997) *J. Biol. Chem.* **272**, 15587–15590
5. Shingledecker, K., Jiang, S. Q., and Paulus, H. (1998) *Gene (Amst.)* **207**, 187–195
6. Mills, K. V., Lew, B. M., Jiang, S.-Q., and Paulus, H. (1998) *Proc. Natl. Acad. Sci. U. S. A.* **95**, 3453–3458
7. Southworth, M. W., Adam, E., Panne, D., Byer, R., Kautz, R., and Perler, F. B. (1998) *EMBO J.* **17**, 918–926
8. Laemmli, U. K. (1970) *Nature* **227**, 680–685
9. Richards, F. M., and Vithayathil, P. J. (1959) *J. Biol. Chem.* **234**, 1459–1464
10. Chaiken, I. M., Freedman, M. H., Lyerla, J. R., and Cohen, J. S. (1973) *J. Biol. Chem.* **248**, 884–891
11. Duan, X., Gimble, F. S., and Quiocho, F. A. (1997) *Cell* **89**, 555–564
12. Reith, M. E., and Munholland, J. (1995) *Plant Mol. Biol. Rep.* **13**, 333–335
13. Perler, F. B., Olsen, G. J., and Adam, E. (1997) *Nucleic Acids Res.* **25**, 1087–1093

EXHIBIT D

Belinda M. Lew[1,2,*]
Kenneth V. Mills[1,2,*]
Henry Paulus[1,3]
[1] Boston Biomedical Research Institute, Boston, MA 02114
[2] Department of Chemistry and Chemical Biology, Harvard University, Cambridge, MA 02138
[3] Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston, MA 02115

Characteristics of Protein Splicing in *trans* Mediated by a Semisynthetic Split Intein

ABSTRACT: *Protein splicing in trans results in the ligation of two protein or peptide segments linked to appropriate intein fragments. We have characterized the trans-splicing reaction mediated by a naturally expressed, approximately 100-residue N-terminal fragment of the Mycobacterium tuberculosis intein and a synthetic peptide containing the 38 C-terminal intein residues, and found that the splicing reaction was very versatile and robust. The efficiency of splicing was nearly independent of temperature between 4 and 37°C and pH between 6.0 and 7.5, with only a slight decline at pH values as high as 8.5. In addition, there was considerable flexibility in the choice of the C-terminal intein fragment, no significant difference in protein ligation efficiency being observed between reactions utilizing the N-terminal fragment and either the naturally expressed 107-residue C-terminal portion of the intein, much smaller synthetic peptides, or the 107-residue C-terminal intein fragment modified by fusion of a maltose binding protein domain to its N-terminus. The ability to use different types of the C-terminal intein fragments and a broad range of reaction conditions make protein splicing in trans a versatile tool for protein ligation.* © 2000 John Wiley & Sons, Inc. Biopoly 51: 355–362, 1999

INTRODUCTION

Protein splicing is a posttranslational processing mechanism that involves the self-catalyzed excision of an intervening sequence, the intein, from flanking polypeptides, the exteins, followed by the ligation of the exteins to form a new peptide bond. The chemical mechanism of protein splicing is well understood and was recently reviewed.[1]

Protein splicing in *trans* was first demonstrated in vivo with the *Mycobacterium tuberculosis* RecA in-

* BML and KVM should be considered joint first authors.
Correspondence to: Henry Paulus, Boston Biomedical Research Institute, 20 Staniford Street, Boston, MA 02114, USA; email: paulus@bbri.org
Contract grant sponsor: National Institute of General Medical Sciences (NIGMS), Howard Hughes Medical Institute, National Institutes of Health (NIH), and National Science Foundation (NSF)
Contract grant number: R01 GM55875 (NIGMS), RR11301 (NIH), and 96-04781 (NSF)
Biopolymers (Peptide Science), Vol. 51, 355–362 (1999)

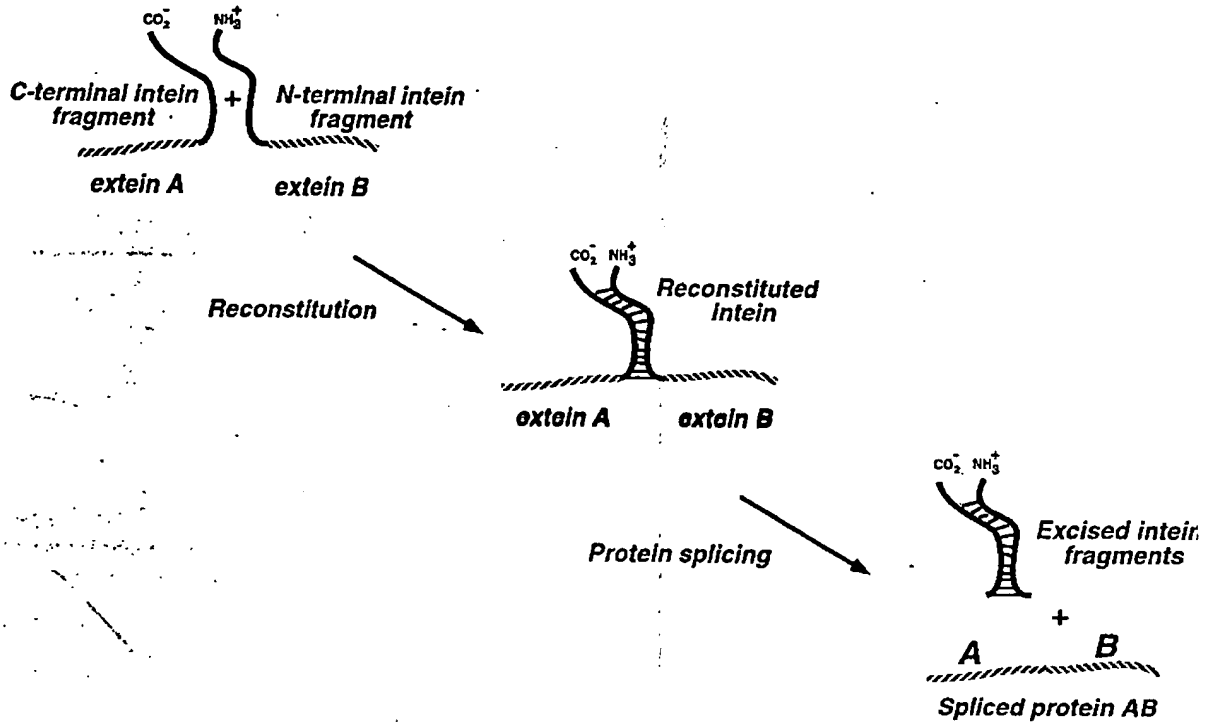


FIGURE 1 The ligation of recombinant proteins or peptides fused to intein fragments mediated by protein *trans*-splicing.

tein[2] and in vitro with the *Pyrococcus furiosus* RIR1 intein[3] and the *Pyrococcus* sp. GB-D Pol1 inteins.[4] We developed an in vitro *trans*-splicing system using separately expressed N- and C-terminal fragments of the *M. tuberculosis* RecA intein, with maltose binding protein and a polypeptide containing a hexahistidine sequence (His-tag) serving as the N- and C-terminal exteins, respectively.[5] Upon combining the denatured RecA intein fragments and renaturation by dialysis under mildly oxidizing conditions, a disulfide-linked dimer could be identified. This dimer could be induced to splice with high efficiency by reduction of the disulfide bond. Our *trans*-splicing system in effect constitutes a protein ligase that promotes the ligation of polypeptides fused to the N- and C-termini of the N- and C-terminal fragments, respectively, of the *M. tuberculosis* RecA intein (Figure 1).

With the aim of providing guidelines for the use of protein splicing as a biochemical-tool, we characterized the conditions under which *trans*-splicing can occur, using a semisynthetic protein splicing system in which a 40-residue peptide ($U_{C38}CA$—a synthetic peptide consisting of the 38 C-terminal residues of U linked to Cys–Ala–NH_2) served as the C-terminal intein fragment plus a dipeptide as the C-extein. In addition, we compared the effectiveness of protein *trans*-splicing using the naturally expressed 107-residue C-terminal intein, the shorter, synthetic C-terminal intein fragment, and the C-terminal intein fragment extended at its N-terminus by fusion to a maltose binding protein domain. The results show protein *trans*-splicing involving fragments of the *M. tuberculosis* RecA intein to be an efficient and versatile reaction with properties suitable for use in protein ligation.

EXPERIMENTAL

Plasmid Constructs and Protein Expression

The construction of plasmid pMU2s/sΔ6, which encodes MU_{NA}, an in-frame fusion of *Escherichia coli* maltose binding protein (MBP) to the 105 N-terminal amino acids of the *M. tuberculosis* RecA intein, followed by the C-terminal sequence Arg–Gly–Glu–Phe, was described earlier.[5] Plasmid pETUH4 encoding $U_{CA}H$, which consists of the N-terminal sequence formyl–Met–Asp–Pro–Ser–Ser–Arg–Ser followed by the 107 C-terminal amino acids of the *M. tuberculosis* RecA intein fused to a 49-amino acid polypeptide with a C-terminal His-tag, was derived from pTrcUH (Ref. 5) by transferring the *NcoI–HindIII* restriction fragment encoding $U_{CA}H$ into pET28a (Novagen, Madison WI). Plasmid pM4DUH, which encodes an in-frame fusion of MBP to the N-terminus of $U_{CA}H$ ($MU_{CA}H$), was constructed from pMU2H (Ref. 2) by replacing a 1-kb segment

extending from the *Aff*II site near the C-terminus of MBP to the *Pst*I site at residue 335 of the intein with an oligonucleotide cassette composed of 5′-TTA AGC TTG GAA GTG CTG TTT CAA GGT CCT GCA and the appropriate complement.

The chimeric proteins encoded by these plasmids were expressed in *Escherichia coli* DH5α, except for $U_{C\Delta}H$, which was expressed in *E. coli* JM109(DE3) (Promega, Madison, WI), as described previously.[5] The cells were disrupted by passage through a French pressure cell after resuspension in buffer O (20 m*M* sodium phosphate, pH 7.5; 500 m*M* NaCl; 1 m*M* EDTA), buffer G (20 m*M* Tris · HCl, pH 7.4; 150 m*M* NaCl; 6*M* GdmCl), or buffer T (20 m*M* Tris · HCl, pH 7.4; 150 m*M* NaCl), for $MU_{N\Delta}$, $U_{C\Delta}H$, or $MU_{C\Delta}H$ (chimeric protein of M fused to $U_{C\Delta}H$), respectively, and the resulting lysates were centrifuged at 15,000× *g* for 35 min. For $MU_{N\Delta}$, the lysate supernatant was passed through a 0.5 mL amylose column (New England Biolabs, Beverly, MA) preequilibrated with buffer O. The column was washed with 15 mL of buffer O and eluted with buffer O supplemented with 10 m*M* maltose. For $U_{C\Delta}H$, the lysate supernatant was purified by batchwise passage through a Talon spin column (Clontech, Palo Alto, CA). The column was washed 2 times with 1 mL of buffer G, 2 times with 1 mL of buffer G supplemented with 10 m*M* imidazole, and eluted with 1 mL buffer G supplemented with 100 m*M* imidazole. $MU_{C\Delta}H$ was purified by the same procedure, except that buffer T was used instead of buffer G.

Protein concentration was determined by the Bradford method[6] or by measuring the absorbance at 280 nm. The molar absorption coefficients for $U_{C38}CA$ and $MU_{C\Delta}H$, 1400 and 74,370 cm^{-1} M^{-1}, respectively, were calculated from their amino acid composition using the PROTEAN program (DNAStar, Madison, WI).

Peptide Synthesis

The 40-residue peptide $U_{C38}CA$ consists of the sequence Glu–Leu–Arg–Tyr–Ser–Val–Ile–Arg–Glu–Val–Leu–Pro–Thr–Arg–Arg–Ala–Arg–Thr–Phe–Asp–Leu–Glu–Val–Glu–Glu–Leu–His–Thr–Leu–Val–Ala–Glu–Gly–Val–Val–Val–His–Asn–Cys–Ala–NH_2. This peptide, which comprises the 38 C-terminal residues of the *M. tuberculosis* RecA intein followed by Cys–Ala–NH_2, was synthesized using N-(9-fluorenyl)methoxycarbonyl chemistry on a peptide synthesizer (Applied Biosystems model 431), utilizing 4-methyl benzhydrylamine resin (Novabiochem, San Diego, CA), *O*-trityl-protected Thr, double coupling of all Thr and Arg residues, an acetic anhydride blocking step at each cycle, and *N*-methylpyrrolidone supplemented with dimethylsulfoxide to 10%. The peptide was cleaved from the resin and deprotected using standard procedures.[7] The peptides were purified by high performance liquid chromatography (HPLC) (Rainin, Woburn, MA) using a preparative C8 column (Vydac, Hesperia, CA). The fractions were monitored by matrix-assisted laser desorption ionization–time of flight (MALDI-TOF) mass spectrometry.

Reassociation and Splicing Conditions

For experiments with the semisynthetic system, 8 μ*M* $MU_{N\Delta}$ was combined with 40 μ*M* $U_{C38}CA$. The mixtures (200 μL) were dialyzed using Spectrapor 1000 MWCO dialysis tubing against 50 mL of buffer M [20 m*M* sodium phosphate, pH 7.5; 500 m*M* NaCl; 1 m*M* EDTA; 1 m*M* tris(2-carboxyethyl)phosphine (TCEP); 8*M* urea] for 1 h, then 3 times against 50 mL of buffer OR (buffer O supplemented with 1 m*M* TCEP) for 20 min each. Following dialysis, which was done at 4°C, the mixtures were incubated at 30°C for appropriate time periods.

The reassociation of $MU_{N\Delta}$ with $U_{C\Delta}H$ was studied by examining the effect of $U_{C\Delta}H$ concentration on the efficiency of the splicing reaction with $MU_{N\Delta}$. $MU_{N\Delta}$ (8 μ*M*) was mixed with appropriate concentrations of $U_{C\Delta}H$ and the mixture (200 μL) was dialyzed against 50 mL buffer N (20 m*M* sodium phosphate, pH 7.5; 500 m*M* NaCl; 1 m*M* EDTA; 1 m*M* TCEP; 6*M* GdmCl) using Spectrapor 8000 MWCO dialysis tubing for 1 h, then twice against 50 mL of buffer OR for 20 min each, at 4°C. This was followed by dialysis against buffer OR at 30°C for 22 h. The reassociation of $MU_{N\Delta}$ with $MU_{C\Delta}H$ was studied in a similar manner except that 10 μ*M* $MU_{N\Delta}$ was used and splicing was allowed to proceed at 25°C for 16 h.

Analysis of Protein Splicing

Sodium dodecyl sulfate–polyacrylamide gel electrophoresis (SDS-PAGE) analysis and biotinylation of free thiols with 1-biotinamido-4-[4′-(maleimidomethyl)-cyclohexanecarboxamido]butane (Biotin-BMCC; Pierce, Rockford, IL), followed by Western blotting and development with streptavidin-linked alkaline phosphatase were performed as described previously.[7] The stained gels and Western blots were scanned with a Supervista S-12 scanner (Umax Technologies, Fremont, CA) and analyzed using the NIH Image 1.60 program. MALDI-TOF mass spectrometry was performed on a Voyager RP Biospectrometry Workstation (PerSeptive Biosystems, Framingham, MA) as described.[7]

RESULTS

The In Vitro *trans*-Splicing System

The products of protein splicing in *trans* are the exteins linked by a normal peptide bond and the excised intein fragments. For example, the spliced product observed as a result of the reaction between $MU_{N\Delta}$ and $U_{C\Delta}H$ is MH.[5] The product of the splicing reaction between $MU_{N\Delta}$ and $U_{C38}CA$ is M linked to Cys–Ala–NH_2 (MCA), i.e., the N-extein, M, linked to the C-extein, Cys–Ala–NH_2 (Figure 2a). Since M does not contain any Cys residues, it was possible to confirm the identity of MCA by biotinylating the reaction products with the cysteine-specific reagent,

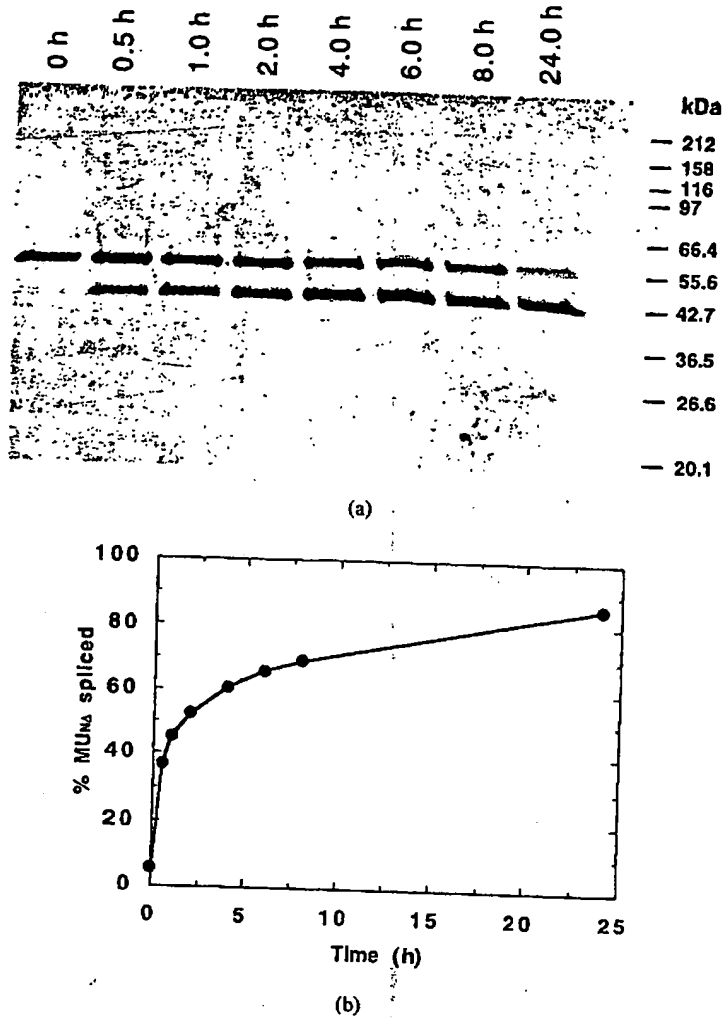


FIGURE 2 Time course of the protein *trans*-splicing reaction. $MU_{N\Delta}$ and $U_{C38}CA$ were reconstituted by renaturation from 8*M* urea at 4°C and pH 7.5, then incubated for various times as described under "Experimental." (a) Aliquots were removed at the times indicated and protein splicing was analyzed by SDS-PAGE. (b) The percent of $MU_{N\Delta}$ converted to spliced product is shown as a function of time.

Biotin-BMCC, followed by Western blotting and reaction with a streptavidin–alkaline phosphatase conjugate.[7] The putative spliced product of M_r of about 43,000 was found to be labeled with biotin, indicating that it consisted primarily of MCA and not M. In control experiments with free M, no labeling was observed under identical conditions (data not shown).

When reassociation was performed in the absence of reductant, disulfide-linked dimers of $MU_{N\Delta}$ with either $MU_{C\Delta}H$ or $U_{C38}CA$ were observed by SDS-PAGE and MALDI-TOF mass spectroscopy, and could be induced to splice by the addition of a disulfide-reducing agent (data not shown). Similar disulfide-linked dimers were described earlier in *trans*-splicing reactions involving $MU_{N\Delta}$ and either $U_{C\Delta}H$ or a 52-residue peptide ($U_{C50}CA$), which contains t 50 C-terminal intein residues plus Cys_T-Ala–NH_2 the C-extein. In the experiments described belo reassociation was performed under reducing conc tions, so that protein splicing occurred without pri formation of disulfide-linked dimers.

Time Course of Protein Splicing

The rate of the protein *trans*-splicing reaction wit $MU_{N\Delta}$ and $U_{C38}CA$ was studied at pH 7.5 and 30° by measuring the conversion of $MU_{N\Delta}$ to splice product by SDS-PAGE (Figure 2a). A small amour of splicing (4–7%) was observed during the reasso ciation step. Upon subsequent incubation at 30°C

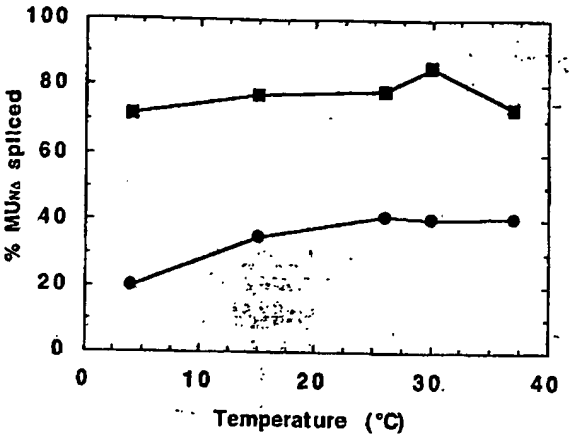


FIGURE 3 Temperature dependence of the protein *trans*-splicing reaction. $MU_{N\Delta}$ and $U_{C38}CA$ were reconstituted from 8*M* urea at 4°C and pH 7.5 as described under "Experimental." Aliquots of the reassociation mixtures were then incubated at the temperatures indicated for 1 h (●) or 24 h (■), and analyzed as described under "Experimental."

50% splicing was attained in 2 h and 86% in 24 h (Figure 2b). The reaction consisted of an initial rapid phase, followed by a slow phase with steadily increasing amounts of spliced product, and could not be described by strict first-order kinetics. For the purpose of comparison in subsequent experiments, the rate and extent of splicing will be defined operationally by the amount of splicing observed after 1 and 24 h, respectively.

Effect of Temperature on Protein Splicing

The *trans*-splicing reaction between $MU_{N\Delta}$ and $U_{C38}CA$ was relatively insensitive to temperature. At 4°C, the rate of splicing was 50% of that at 15°C; further increase of temperature to 37°C had little additional effect (Figure 3). The extent of the *trans*-splicing was about 70% in the entire temperature range studied, with a maximum observed value of 85% at 30°C.

Effect of pH on Protein Splicing

The rate and extent of splicing between $MU_{N\Delta}$ and $U_{C38}CA$ were measured as a function of pH between 6.0 and 8.2 at 30°C. The rate of splicing was about 3 times greater at pH 6.0 than at pH 8.2, but its extent was relatively insensitive to pH differences between 6.0 and 7.5, with a slight decline at higher pH values (Figure 4).

Effect of Shortening and Lengthening of the C-Terminal Intein Fragment on Protein Splicing Efficiency

The extent of conversion of a constant amount of $MU_{N\Delta}$ to spliced product when incubated with varying amounts of a C-terminal intein fragment was compared using different types of C-terminal fragments. In contrast to an earlier experiment of this type,[7] a very low concentration of $MU_{N\Delta}$ (8 μM) was used in these experiments to make the interaction of the N- and C-terminal intein fragments a limiting factor. Protein splicing was allowed to occur at conditions under which the extent of protein splicing is expected to reach its maximum value. When 8 μM $MU_{N\Delta}$ was reconstituted with various concentrations of the naturally expressed $U_{C\Delta}H$, which includes the 107 C-terminal intein residues, a maximum conversion of about 85% $MU_{N\Delta}$ to spliced product was reached at about a 1:10 molar ratio, with 60% yield at equimolar amounts of $MU_{N\Delta}$ and $U_{C\Delta}H$ (Figure 5a).

The splicing efficiency of a highly truncated C-terminal intein fragment was examined by reconstituting 8 μM $MU_{N\Delta}$ with varying concentrations of $U_{C38}CA$, which contains only the 38 C-terminal residues of the intein. Maximum splicing yield (about 75%) was achieved with a 5-fold molar excess of $U_{C38}CA$, with 48% yield at equimolar amounts of $MU_{N\Delta}$ and $U_{C38}CA$ (Figure 5b). This contrasts with our earlier observation that splicing with $U_{C50}CA$, which contains the 50 C-terminal intein residues, reached its maximum yield at a 1:1 molar ratio of

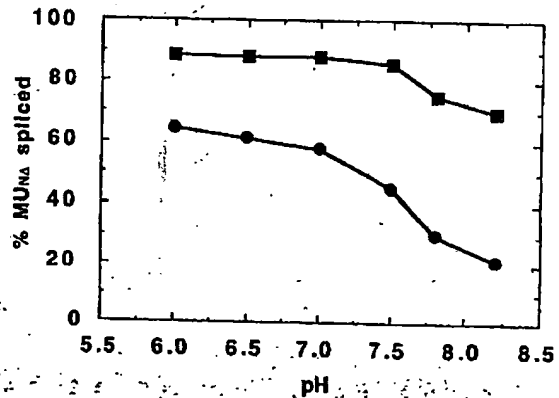
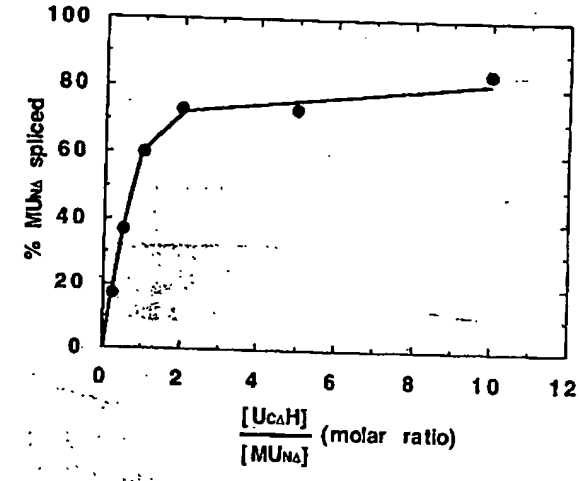
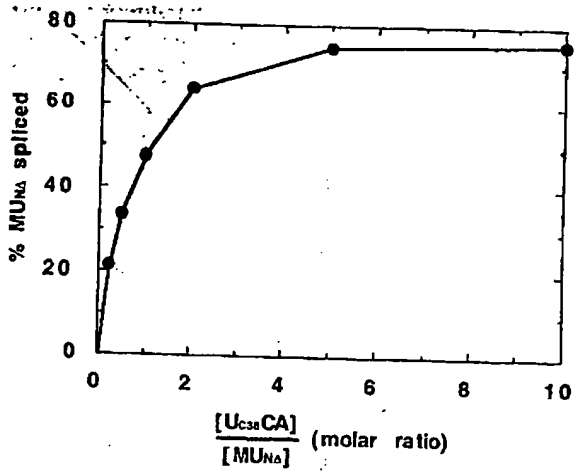


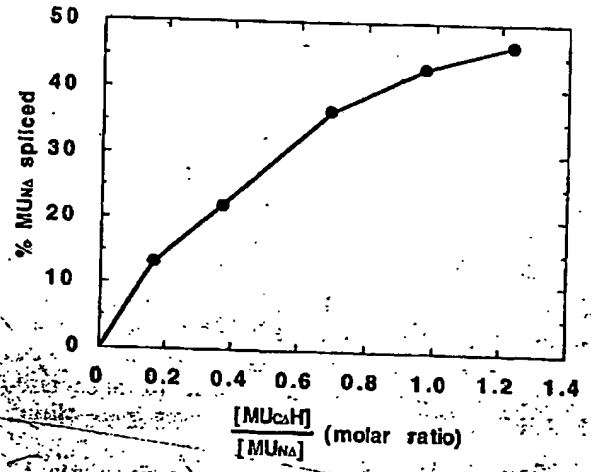
FIGURE 4 Effect of pH on the protein *trans*-splicing reaction. $MU_{N\Delta}$ and $U_{C38}CA$ were reconstituted from 8*M* urea at 4°C and pH 7.5 as described under "Experimental." Samples of the reassociation mixtures were adjusted to the pH values indicated by mixing with an equal volume of 200 m*M* sodium phosphate buffer of the appropriate pH value. The fraction of $MU_{N\Delta}$ converted to spliced product after 1 h (●) or 24 h (■) was determined as described under "Experimental."



(a)



(b)



(c)

FIGURE 5 · Effect of C-terminal fragment length and concentration on percent yield of spliced product. Various concentrations of (a) $U_{C\Delta}H$, (b) $U_{C38}CA$, or (c) $MU_{C\Delta}H$ were reconstituted with a constant amount of $MU_{N\Delta}$ (8.0 μM with $U_{C\Delta}H$ or $U_{C38}CA$; 10 μM with $MU_{C\Delta}H$) as described under "Experimental." Following splicing for 22, 30, or 16 h, respectively, at 30°C, the samples were analyzed for the extent of protein splicing as described in "Experimental." The extent of conversion of $MU_{N\Delta}$ to spliced product is shown as a function of the molar ratio of the C-terminal fragment to $MU_{N\Delta}$.

$MU_{N\Delta}$ and the C-terminal fragment.[7] However should be noted that this earlier experiment was formed at 42 μM $MU_{N\Delta}$. When repeated with 8 $MU_{N\Delta}$, the maximum splicing yield (69%) achieved only when using a 5-fold molar excess of C-terminal intein fragment (data not shown).

In order to determine whether fusing a large p tein to the N-terminus of the C-terminal intein fr ment compromises its ability to function in prot splicing, we examined $MU_{C\Delta}H$, an in-frame fusion MBP with $U_{C\Delta}H$. As shown in Figure 5c, differ concentrations of $MU_{C\Delta}H$ promoted the splicing $MU_{N\Delta}$, almost to the same extent as $U_{C38}CA$ (Fig 5b), with 44% splicing at a 1:1 molar ratio. It sho be noted that the concentration of $MU_{N\Delta}$ was sligh higher in this experiment (10 μM) and that the limit amount of $MU_{C\Delta}H$ available did not allow the det mination of splicing efficiency at saturating conce trations of the C-terminal intein fragment.

DISCUSSION

In the period between its discovery in 1990 and abo 1996, the study of protein splicing focused primari on its mechanism. The elucidation of the chemic reactions that underlie protein splicing[1] has opene the way for harnessing protein splicing for prote engineering.[8] Of special interest in this connection the question of whether protein splicing elements c serve as efficient tools for protein ligation.

A purely chemical approach to polypeptide lig tion has been developed in the laboratories of Ken and Tam.[10] One of the peptides to be ligated is sy thesized as a C-terminal thioester derivative; the oth peptide has a N-terminal cysteine residue. The *tran* thioesterification is initiated by attack of the cysteiny thiol on the thioester, resulting in a new protein wit a thioester linkage between the two fragments. Thi thioester rapidly converts to a peptide bond via a S–N-acyl rearrangement, analogous to the last step o protein splicing.[11] This method, sometimes referred t as "native chemical ligation," has been used for th synthesis of small polypeptides in yields of nearl 90%[10] as well as of small proteins, such as the 72 residue IL-8,[9] but is constrained by the requiremen

that the N-terminal fragment be synthesized with a C-terminal thioester and by the practical size limit for peptide synthesis.

The laboratories of Muir[12,13] and Xu[14] have made use of a protein splicing element to produce protein thioester derivatives, which then can undergo ligation with a polypeptide that has a N-terminal cysteine, as in native chemical ligation. This method uses a modified intein that can catalyze only the first two steps of protein splicing, N–S-acyl rearrangement and *trans*-esterification, to accumulate intermediates in which the C-terminus of the N-extein is a reactive thioester.[15] When thiols such as thiophenol or mercaptoethanesulfonic acid and a synthetic peptide with a N-terminal cysteine are added in a large excess, a series of *trans*-esterifications followed by an S–N-acyl rearrangement leads to the ligation of the N-extein with the synthetic peptide. This method, which has been referred to as "expressed protein ligation" or "protein semisynthesis," differs from native chemical ligation in that it does not rely solely on synthetic peptides but allows the ligation of any protein that can be expressed in *E. coli* with a synthetic peptide. It has been used for the semisynthesis of transcription factors,[13] protein tyrosine kinase,[12] ribonuclease A, and a restriction endonuclease.[14] However, a limitation of this approach is the need to use a large excess of the synthetic peptide that is to be ligated.

Protein *trans*-splicing harnesses the power of the complete splicing reaction, not only the first step, and could potentially eliminate many of the limitations of the other approaches to protein ligation. Yamazaki and co-workers[3] used the *P. furiosus* RIR1 intein for selectively labeling the N-terminal portion of the α subunit of *E. coli* RNA polymerase with ^{15}N, albeit in very low yield. Perler and co-workers[4] demonstrated protein *trans*-splicing in vitro using the *Psp* Pol-1 intein, but were unable to retain function after deleting the homing endonuclease domain. In contrast, the *M. tuberculosis* RecA intein has been more amenable to truncation, retaining the ability to mediate efficient *trans*-splicing in vitro not only after the elimination of the entire homing endonuclease domain[5] but also after substituting synthetic peptides 35–50 residues in length for the C-terminal intein fragment.[7]

Our results show protein *trans*-splicing mediated by fragments of the *M. tuberculosis* RecA intein to be a very versatile and robust reaction. The splicing efficiency is nearly independent of temperature between 4 and 37°C (Figure 3) and pH between 6.0 and 7.5, with only a modest decline up to pH 8.5 (Figure 4). Accordingly, it is possible to choose conditions for protein ligation that are most compatible with the stability of the target protein. In addition, there is considerable flexibility in the choice of the C-terminal intein fragment, which can either be the naturally expressed 107-residue C-terminal portion of the intein, a much smaller synthetic peptide, comprising as few as 38 of the C-terminal intein residues, or the 107-residue C-terminal intein fragment modified by fusing an affinity tag as large as the 43-kDa MBP to its N-terminus. No significant differences in protein ligation efficiencies were noted when any of these were reconstituted with the 105-residue N-terminal intein fragment (Figure 5). An advantage of using a synthetic peptide as the C-terminal intein fragment is that it can be synthesized together with the C-extein. This provides the opportunity for introducing specifically labeled or unnatural amino acids into the C-terminal portion of the ligated protein as probes for studying its structure or function. An advantage of using an expressed C-terminal intein fragment linked to an affinity tag is the ease of rapid purification under mild conditions, regardless of the C-extein to which it is fused, which is especially important if the protein to be ligated is relatively unstable.

A fundamental difference in the application of expressed protein ligation and protein *trans*-splicing to protein ligation lies in the effective molecularities of these reactions. Expressed protein ligation is a strict bimolecular reaction. In most cases, the two reactants have no affinity for each other and efficient ligation of the protein component to the synthetic polypeptide needs a substantial excess of the latter, which is ordinarily used at millimolar concentrations.[12–14] In contrast, the first step of protein *trans*-splicing is the formation of a complex of the two intein fragments, which occurs with relatively high affinity, as evidenced by the almost quantitative production of disulfide-linked dimers when fusion proteins with the N- and C-terminal fragments of the *M. tuberculosis* RecA intein are mixed in the micromolar concentration range.[5] Once the complex is formed, protein ligation is essentially an intramolecular reaction, and as a result, about 50% conversion of 8 μM $MU_{N\Delta}$ to spliced product is achieved with an equimolar concentration and 70–80% conversion with a 5-fold excess (40 μM) of the C-terminal intein fragment (Figure 5). On the other hand, the need for prior complex formation in protein *trans*-splicing has the potential disadvantage that the formation of such complexes is achieved most efficiently by prior denaturation, followed by renaturation.[5,7] This limits the application of protein *trans*-splicing to proteins that can be reversibly denatured. However, protein ligation in the semisynthetic *trans*-splicing system can also occur, albeit at a slower rate, without prior denaturation (BML and KVM, unpublished observa-

tions), allowing ligations involving proteins whose denaturation is irreversible.

In expressed protein ligation, the C-terminal protein fragment to be ligated has to be available in large amounts for use at millimolar concentrations, and is therefore usually a synthetic polypeptide. Since solid-phase peptide synthesis becomes relatively inefficient for peptides with more than 50–70 amino acids, this imposes a size limit to the C-terminal moiety of the ligated protein. Very recently, however, methods have become available for generating large polypeptides with N-terminal cysteine residues, either by specific proteolysis of an expressed recombinant protein[16,17] or by using a novel intein-based expression system,[18,19] thus considerably expanding the scope of expressed protein ligation. Protein *trans*-splicing can mediate the ligation of any set of proteins that can be expressed as fusion proteins with the intein fragments, so that there is essentially no size limit on either the N- or the C-terminal moiety of the protein to be ligated. This will give considerable flexibility in expressing toxic proteins as two nontoxic fragments that are subsequently ligated in vitro and will also allow novel types of protein recombination such as post-translational domain swapping of large multidomain proteins.

Besides serving as a tool for protein ligation, the *trans*-splicing system described here and in earlier papers[5,7] will be useful in addressing questions about protein splicing itself. For example, the interaction of the N- and C-terminal intein fragments can be studied directly, or using the semisynthetic protein splicing element, after introducing various probes or non-natural amino acids. Protein *trans*-splicing may also provide some insights into intein biosynthesis. In most inteins, a homing endonuclease domain interrupts the intein. In the course of translation, the N-terminal half of the intein is expressed first, followed by the synthesis of the homing endonuclease region, which folds into an independent globular domain,[20] and finally by the translation of the C-terminal half of the intein, which presumably cannot fold productively independent of the N-terminal half. That inteins can be reconstituted after denaturation suggests that intein folding is not strictly cotranslational but that the N-terminal half of the protein splicing domain may act as a chaperone for assuring the proper folding of the C-terminal portion.

This investigation was supported by grant R01 GM558 from the National Institute of General Medical Scien (HP) and by a Howard Hughes Medical Institute Pred toral Fellowship (KVM). The mass spectrometer used this work was funded by grants from the National Institu of Health (RR11301) and the National Science Foundati (96-04781). We thank Paul Morgan for help and advi Shu-qin Jiang and Paul Morgan for constructing the pl mids, Anna Wong for synthesizing the peptides, and Pa Leavis and Betty Gowell for use of their preparative HPL

REFERENCES

1. Paulus, H. Chem Soc Rev 1998, 27, 375–386.
2. Shingledecker, K.; Jiang, S.-Q.; Paulus, H. Gene 199 207, 187–195.
3. Yamazaki, T.; Otomo, T.; Oda, N.; Kyogoku, Y.; U gaki, K.; Ito, N.; Ishino, Y.; Nakamura, H. J Am Che Soc 1998, 120, 5591–5592.
4. Southworth, M. W.; Adam, E.; Panne, D.; Byer, R Kautz, R.; Perler, F. B. EMBO J 1998, 17, 918–926
5. Mills, K. V.; Lew, B. M.; Jiang, S.-Q.; Paulus, H. Pro Natl Acad Sci USA 1998, 95, 3453–3458.
6. Bradford, M. M. Anal Biochem 1976, 72, 248–254.
7. Lew, B. M.; Mills, K. V.; Paulus, H. J Biol Chem 199 273, 15887–15890.
8. Gimble, F. S. Chem Biol 1998, 5, R251–R256.
9. Dawson, P. E.; Muir, T. W.; Clark-Lewis, I.; Ken S. B. H. Science 1994, 266, 776–779.
10. Tam, J. P.; Lu, Y.-A.; Liu, C.-F.; Shao, J Proc Na Acad Sci USA 1995, 92, 21485–12489.
11. Shao, Y.; Paulus, H. J Peptide Res 1997, 50, 193–19
12. Muir, T. W.; Sondhi, D.; Cole, P. A. Proc Natl Acad Sc USA 1998, 95, 6705–6710.
13. Severinov, K.; Muir, T. W. J Biol Chem 1998, 273 16205–16209.
14. Evans, T. C.; Benner, J.; Xu, M.-Q. Protein Sci 1998, 7 2256–2264.
15. Chong, S.; Shao, Y.; Paulus, H.; Benner, J.; Perler F. B.; Xu, M. Q. J Biol Chem 1996, 271, 22159–22168
16. Cotton, G. J.; Ayers, B.; Xu, R.; Muir, T. W. J Am Chem Soc 1999, 121, 1100–1101.
17. Xu, R.; Ayers, B.; Cowburn, D.; Muir, T. W. Proc Natl Acad Sci USA 1999, 96, 388–393.
18. Evans, T. C.; Benner, J.; Xu, M.-Q. J Biol Chem 1999, 274, 3923–3926.
19. Mathys, S.; Evans, T. C.; Chute, I. C.; Wu, H.; Chong, S.; Benner, J.; Liu, X.-Q.; Xu, M.-Q. Gene 1999, 231, 1–13.
20. Duan, X.; Gimble, F. S.; Quiocho, F. A. Cell 1997, 89, 555–564.

Exhibit E

Proc. Natl. Acad. Sci. USA
Vol. 95, pp. 9226–9231, August 1998
Biochemistry

Protein *trans*-splicing by a split intein encoded in a split DnaE gene of *Synechocystis* sp. PCC6803

HONG WU, ZHUMA HU, AND XIANG-QIN LIU*

Biochemistry Department, Dalhousie University, Halifax, Nova Scotia, B3H 4H7, Canada

Edited by Allan Campbell, Stanford University, Stanford, CA, and approved June 9, 1998 (received for review April 6, 1998)

ABSTRACT **A split intein capable of protein *trans*-splicing is identified in a DnaE protein of the cyanobacterium *Synechocystis* sp. strain PCC6803. The N- and C-terminal halves of DnaE (catalytic subunit α of DNA polymerase III) are encoded by two separate genes, *dnaE-n* and *dnaE-c*, respectively. These two genes are located 745,226 bp apart in the genome and on opposite DNA strands. The *dnaE-n* product consists of a N-extein sequence followed by a 123-aa intein sequence, whereas the *dnaE-c* product consists of a 36-aa intein sequence followed by a C-extein sequence. The N- and C-extein sequences together reconstitute a complete DnaE sequence that is interrupted by the intein sequences inside the β- and τ-binding domains. The two intein sequences together reconstitute a split mini-intein that not only has intein-like sequence features but also exhibited protein *trans*-splicing activity when tested in *Escherichia coli* cells.**

Inteins have been defined as protein sequences embedded in-frame within a precursor protein sequence and excised during a maturation process termed protein splicing (1, 2). Protein splicing is a post-translational event involving precise excision of the intein sequence and concomitant ligation of the flanking sequences (N- and C-exteins) by a normal peptide bond (3–5). Most reported inteins are thought to be bifunctional elements, possessing a protein splicing activity and an endonuclease activity (6–9). Crystal structure of the *Sce* VMA1 intein revealed a two-domain structure, with domain I consisting of the N- and C-terminal regions of the intein sequence and domain II formed by the middle part of the intein sequence (10). Domain I (or a part of it) was suggested to be the splicing domain, whereas domain II corresponded to the endonuclease domain. Such a bipartite structure may be applicable to many other inteins, as has been suggested by studies including mutagenesis (11, 12) and sequence statistical modeling (7–9). Functional studies of mini-inteins, either found in nature or engineered *in vitro*, also confirmed such a two-domain model (13–15), further suggesting that the N- and C-terminal regions of an intein make up a functional splicing domain. Molecular mechanisms of protein splicing involve an N→S (or N→O) acyl shift at the N-terminal splice site (16–18), formation of a branched intermediate (19, 20), and cyclization of an invariant Asn residue at the C terminus of intein to form succinimide (21), leading to excision of the intein. The ligated exteins undergo an S→N (or O→N) acyl shift to form a native peptide bond (21). Amino acid residues that are implicated in the splicing mechanism include a nucleophilic amino acid (Cys, Ser, or Thr) both at the beginning of the intein sequence and at the beginning of the C-extein sequence, an internal His, and a His–Asn dipeptide at the end of the intein sequence. In crystal structures of two inteins, these amino acids are indeed positioned at or near the active site of protein splicing (10, 22).

Approximately 50 intein-coding sequences have been found in >20 different genes distributed among the nuclear and organellar genomes of eukaryotes, archaebacteria (archaea), and eubacteria, suggesting a wide distribution of inteins (see the Intein Registry at http://www.neb.com/neb/inteins.html). Inteins, like many introns (23), are thought to be mobile genetic elements that can be transmitted through horizontal transfer (intein homing), and the intein endonuclease activity is thought to initiate this process (24–26). Known inteins share little overall sequence identity, except between homologous inteins found at the same insertion site in homologous proteins of different organisms (6). A number of short sequence motifs do show a low but significant degree of conservation among inteins (6, 27), suggesting similarities in intein structure, function, and evolutionary origin. Previously reported inteins all have continuous sequences, most are 400–500 aa in size with a protein splicing domain and an endonuclease domain, whereas a few mini-inteins are ≈150 aa in size with a splicing domain only. Three intein sequences were found previously in the cyanobacterium *Synechocystis* sp. strain PCC6803 (*Ssp*), including the *Ssp* DnaB intein in a DNA helicase (28), the *Ssp* DnaX intein in the τ subunit of DNA polymerase III (29), and the *Ssp* GyrB intein in a DNA gyrase B subunit (7, 9). Here, we report a new intein (*Ssp* DnaE intein) found in this cyanobacterium and present in a DnaE protein. DnaE is the catalytic subunit of bacterial DNA polymerase III. In *E. coli*, DNA polymerase III holoenzyme is the replicative polymerase responsible for the synthesis of the majority of the genome. DnaE (also known as α), in addition to its catalytic role, also serves as an organization protein to hold the 18-protein holoenzyme complex together. Its C-terminal half interacts directly with the τ subunit to form a dimeric polymerase and with the β subunit that forms a sliding clamp on the DNA template, whereas its N-terminal half contains the polymerase active site (30). In this study, we show that the DnaE protein of *Synechocystis* sp. PCC6803 is encoded by a split gene interrupted by intein sequences. In an independent study, Gorbalenya also predicted this intein-containing split DnaE gene through sequence analysis (39). We further demonstrate that the products of the split DnaE gene can undergo protein *trans*-splicing to form an intact DnaE protein.

EXPERIMENTAL PROCEDURES

DNA Sequence Analysis and Cloning. The BLAST search program (31) was used in GenBank searches. Protein sequence alignments were produced by using the CLUSTAL W program (32) followed by hand fitting. The *Ssp dnaE*-coding sequences were prepared from total DNA of *Synechocystis* sp. strain PCC6803 (*Ssp*) by a PCR using the thermostable DNA poly-

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

This paper was submitted directly (Track II) to the *Proceedings* office.
Abbreviations: *Ssp*, *Synechocystis* sp. PCC6803; DnaE, catalytic subunit α of DNA polymerase III.
*To whom reprint requests should be addressed. e-mail: pxqliu@is.dal.ca.

merase Pfu (Stratagene). The 2,694-bp *dnaE-n* gene was amplified by using a pair of oligonucleotide primers: 5′-ATGTCCTTCGTCGGTCYTCCATATC-3′ and 5′-ATCAATAAATCGCCTTCACATTGTAATC-3′. The 1,377-bp *dnaE-c* gene was amplified by using a pair of oligonucleotide primers: 5′-ATGGTTAAAGTTATCGGTCGTCGTTC-3′ and 5′-CTAGCCAACACTCTGGCTTTGG-3′. A recombinant expression plasmid was constructed as a tripartite fusion of the complete *dnaE-c* sequence, a portion of the *dnaE-n* sequence (named *dnaE-n′*, 1,017 bp), and the expression plasmid vector pET-32 (Novagen) without the thioredoxin gene. A cassette of termination codon followed by Shine-Dalgarno sequence followed by initiation codon was inserted between the two genes by a PCR-mediated method. First, a linear DNA fragment was amplified from the circular plasmid DNA in a PCR, using the Advantage cDNA polymerase mix (CLONTECH) and a pair of oligonucleotide primers: 5′-TTAATAATAATGGGTACCTTGAAAATGGATTTTTTAGGCTTG-3′, and 5′-ATTATTATTAACCTCCTTAACTCTGGCTTTGGGGTAACAGTGG-3′. The amplified linear DNA molecule was circularized to form the expression plasmid.

Protein Production and Splicing in *E. coli* Cells. The expression plasmid containing *dnaE-c* and *dnaE-n′* sequences was used to transform *E. coli* cells. The transformed cells were grown in liquid Lurie Broth medium at 37°C to late log phase (A_{600}, 0.5). Isopropyl β-D-thiogalactoside (IPTG) was added to a final concentration of 0.8 mM to induce production of the recombinant proteins, and the induction was continued overnight at 15°C. Cells were lysed in SDS-containing loading buffer in a boiling water bath before SDS/PAGE. Antisera used in Western blots were raised in rabbits against specific antigens that had been overproduced in *E. coli* cells transformed with the corresponding genes. The anti-N antiserum was raised against the complete DnaE-n protein. The anti-C antiserum was raised against the first 400 aa of the DnaE-c protein. The specificity of each antiserum was confirmed by testing on the corresponding antigen. The amount of protein in individual protein bands was estimated by using a gel documentation system (Gel Doc 1000 coupled with MOLECULAR ANALYST software, Bio-Rad). A protein band of interest was excised from SDS-polyacrylamide gel after staining, and the protein was electro-eluted and transferred onto poly(vinylidene difluoride) membrane for protein micro-sequencing. In peptide analysis and sequencing, the protein of interest was treated with protease trypsin, the resulting peptides were resolved by HPLC chromatography, peptides of interest were screened by mass spectrometry, and selected peptides were subjected to micro-sequencing. Protein and peptide sequencing, protease digestion, and peptide analysis were all carried out at the Microchemistry Facility of Harvard University.

RESULTS

Sequence Analysis of the Split DnaE Genes. The complete genome sequence has been determined previously for *Synechocystis* sp. PCC6803 (33), and a list of the gene content can be seen at the CyanoBase web site (http://www.kazusa.or.jp/cyano/cyano.html). In browsing through this CyanoBase, we noticed that there are two separate ORFs (ORFs slr0603 and sll1572) showing significant sequence similarities to the *E. coli* DnaE protein (DNA polymerase III α subunit). Further analysis revealed that ORF slr0603 and ORF sll1572 are two members of a discontinuous (split) DnaE gene, and these ORFs subsequently were named *dnaE-n* and *dnaE-c*, respectively (Fig. 1). The *dnaE-n*-coding sequence is 2,694 bp long and spans from base 3,561,946 to 3,564,639 of the genome. The *dnaE-c*-coding sequence is 1,377 bp long and spans from base 737,811 to 736,435 of the genome. These two genes are separated by 745,226 bp of sequence and numerous unrelated

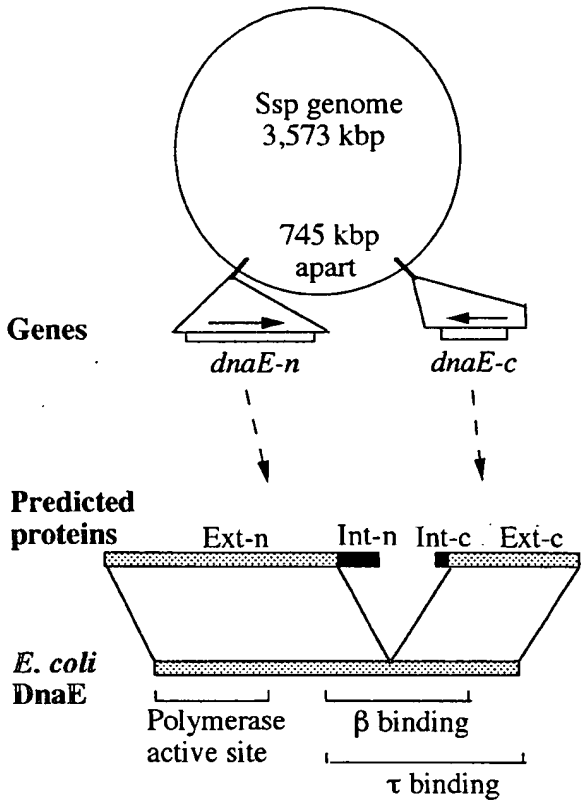


FIG. 1. Gene map and protein structure. Two members of the split DnaE gene, *dnaE-n* and *dnaE-c*, are shown on the genome of *Synechocystis* sp. PCC6803 (*Ssp* genome). In the predicted proteins, DnaE-related sequences (▨) are specified as exteins Ext-n and Ext-c, whereas intein-related sequences (■) are specified as Int-n and Int-c. The exteins are related to *E. coli* DnaE protein whose functional domains are marked.

genes on the 3,573,470-bp circular genome. In addition to distance, coding sequences of these two genes are located on opposite DNA strands. There is no indication of intron sequence either downstream of *dnaE-n* or upstream of *dnaE-c*. In fact, the *dnaE-n* gene is followed immediately downstream by a *lepA* gene that encodes a GTP-binding protein unrelated to DnaE, with a 199-bp intergenic spacer between them. The *dnaE-c* gene is flanked upstream by an unidentified ORF that is unrelated to DnaE and has some similarity to lysostaphin, with a 215-bp intergenic spacer between them. There is no additional DnaE-like gene listed in the CyanoBase. We also were unable to find an additional DnaE gene (complete or in fragments) either by extensive BLAST searches of the complete *Ssp* genome sequence or by Southern blot analysis of the total *Ssp* DNA by using the *Ssp* DnaE gene and the *E. coli* DnaE gene as DNA probes (data not shown).

Protein sequence deduced from the *dnaE-n* gene can be divided into two regions: a 774-aa extein region named Ext-n followed by a 123-aa intein region named Int-n. Similarly, protein sequences deduced from the *dnaE-c* gene can be divided into an intein region (Int-c, 36 aa) followed by an extein region (Ext-c, 423 aa). The Ext-n and Ext-c sequences correspond to the N- and C- terminal halves of a DnaE protein, respectively, and together they reconstitute a complete DnaE sequence. This *Ssp* DnaE sequence, although discontinuous or split, resembles the continuous DnaE sequences of other organisms both in length and in sequence (Fig. 2*A*). The *Ssp* DnaE sequence is 36%, 37%, and 35% identical to DnaE proteins of *E. coli*, *Bacillus subtilis*, and *Mycobacterium tuberculosis*, respectively, over the entire 1,196 aa sequence. These

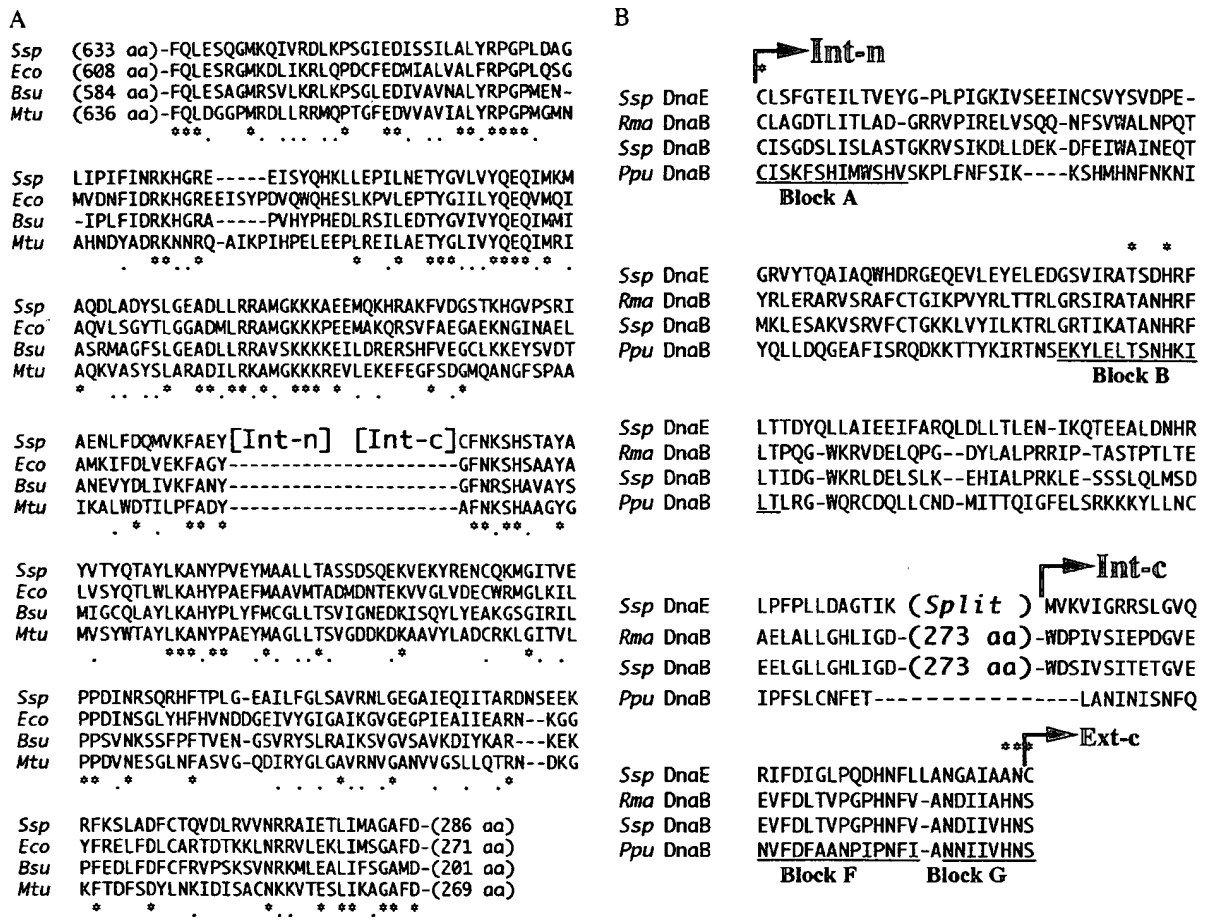


FIG. 2. Sequence analysis. (*A*) Sequence comparison to DnaE proteins. The *Ssp* DnaE extein sequences (*Ssp*) are aligned with corresponding DnaE sequences of *E. coli* (*Eco*), *Bacillus subtilis* (*Bsu*), and *Mycobacterium tuberculosis* (*Mtu*). Only sequences proximal to the intein sequences (Int-n and Int-c) are shown, whereas the number of omitted residues at the N- and C-termini are shown in parentheses. Symbols: – represent gaps introduced to optimize the alignment; * and . mark positions of identical and similar amino acids, respectively. (*B*) Sequence comparison to inteins. The *Ssp* DnaE intein sequences (*Ssp* DnaE), consisting of Int-n and Int-c as indicated, are aligned with corresponding sequences of *Rhodothermus marinas* DnaB intein (*Rma* DnaB), *Synechocystis* sp. PCC6803 DnaB intein (*Ssp* DnaB), and *Porphyra purpurea* chloroplast DnaB intein (*Ppu* DnaB). In the *Rma* DnaB intein and the *Ssp* DnaB intein, only sequences relating to Int-n and Int-c are shown, whereas the number of omitted residues are shown in parentheses. Putative intein motifs (Blocks A, B, F, and G) are underlined, with several critical residues marked by *.

degrees of sequence identity are comparable with the 35–36% sequence identities found among DnaE proteins of the other three compared bacterial organisms.

The Int-n and Int-c sequences show no detectable similarity to DnaE proteins but instead have marked similarity to known intein sequences (Fig. 2*B*). Int-n and Int-c correspond to the N- and C-terminal halves of the intein, and together they reconstitute a mini-intein sequence (named *Ssp* DnaE intein) with a composite length of 159 aa. The sequence of this discontinuous (split) *Ssp* DnaE intein is most similar to corresponding sequences of the *Rma* DnaB intein found previously in a DnaB protein (DNA helicase) of the thermophilic eubacterium *Rhodothermus marinus* (34). The *Ssp* DnaE intein sequence is 30% identical to the *Rma* DnaB intein and 22% identical to the *Ssp* DnaB intein over the 159-aa sequence. Much lower sequence identities were found in comparing it with other known inteins. The *Ssp* DnaE intein, in addition to being split, lacks sequences for a centrally located endonuclease domain that is present in most known inteins including the *Rma* DnaB intein. Nevertheless, the split *Ssp* DnaE intein has many known sequence features of an intein splicing domain. A 50% sequence identity was found between the *Ssp* DnaE intein and the *Rma* DnaB intein over the conserved sequence blocks (A, B, F, and G, totaling 49 aa). Residues important for the catalysis of protein splicing were found in the *Ssp* DnaE intein, including a nucleophilic residue (Cys) at the beginning of the intein sequence, another Cys at the beginning of the C-extein, a Thr and a His in sequence block B, and an Asn at the end of the intein. An Ala precedes the C-terminal Asn in the *Ssp* DnaE intein, although this position is occupied by His in most, but not all, known inteins.

The insertion site of the split *Ssp* DnaE intein is inside the β- and τ-binding domains but outside the polymerase active site of the DnaE protein, according to a comparison with the better studied *E. coli* DnaE protein (Fig. 1). The *Ssp* DnaE intein disrupts a conserved region of the DnaE sequence (Fig. 2*A*), which helped to define the extein–intein boundaries. The first residue of Ext-c in the *Ssp* DnaE sequence is Cys, whereas this position in the other DnaE proteins is occupied by Gly or Ala. This observation is consistent with a requirement of the Cys in *Ssp* DnaE for protein splicing and the absence of an intein in the other DnaE proteins.

Protein *Trans*-Splicing. The split *Ssp* DnaE intein was tested in *E. coli* cells for protein *trans*-splicing activity (Fig. 3). The DnaE-n- and DnaE-c-coding sequences were inserted into an expression plasmid vector to form a two-gene operon (Fig. 3*A*), allowing production of the two proteins inside the same *E. coli* cell and from a single inducible promoter. The construct contained the complete DnaE-c-coding sequence and a partial DnaE-n-coding sequence. Using a complete DnaE-n-coding

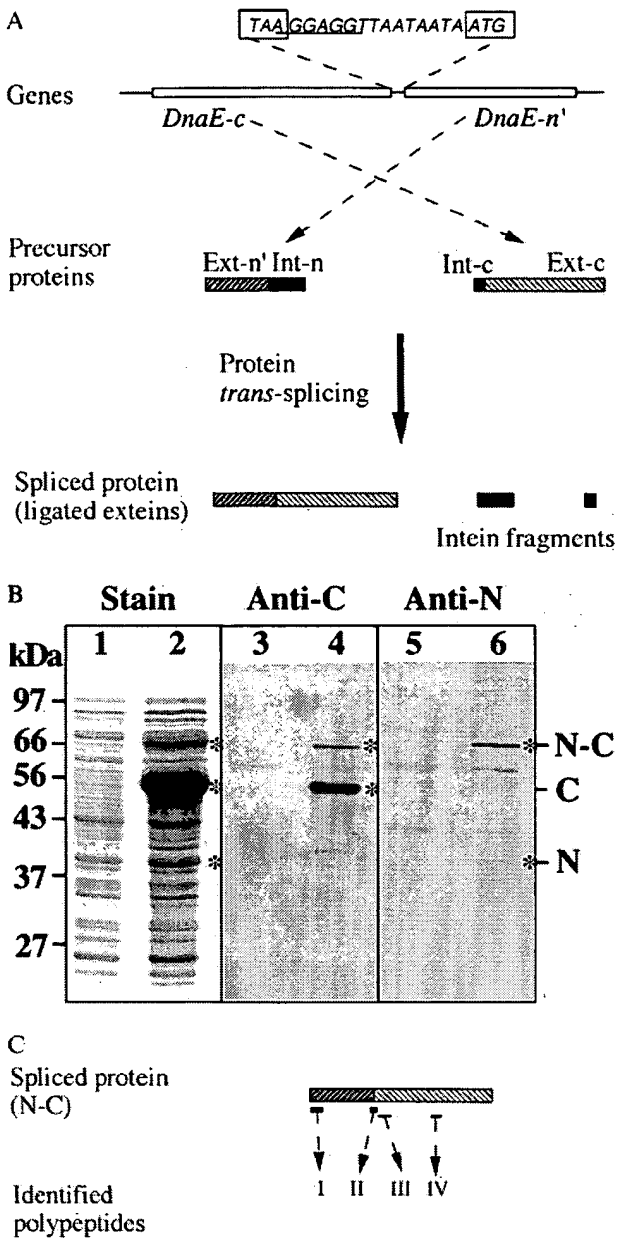


FIG. 3. Protein *trans*-splicing. The *dnaE-n* and *dnaE-c* genes are co-expressed in *E. coli* cells to observe protein *trans*-splicing. (*A*) Schematic illustration. The genes are constructed as a two-gene operon in an expression plasmid vector, with the complete DnaE-c-coding sequence followed by a partial DnaE-n-coding sequence (DnaE-n'). In the intergenic spacer, the termination codon (*TAA*) of DnaE-c and the initiation codon of DnaE-n' are boxed, and the Shine-Dalgarno sequence (ribosome-binding site) is underlined. Products of the two genes are shown as precursor proteins, with their extein regions (Ext-n' and Ext-c) and intein regions (Int-n and Int-c) as indicated. Protein *trans*-splicing produces a spliced protein and excised intein fragments. (*B*) Protein gels. Total proteins of uninduced cells (lanes 1, 3, 5) and induced cells (lanes 2, 4, 6) were resolved by SDS/PAGE and visualized by staining (lanes 1 and 2), by Western blotting with anti-C (DnaE-c) antiserum (lanes 3 and 4), or by Western blotting with anti-N (DnaE-n) antiserum (lanes 5 and 6). Positions of precursor proteins (N and C) and the spliced protein (N-C) are marked. (*C*) Identification of the spliced protein. Peptides I and II were identified by sequencing, and the determined sequences are shown (? marks undetermined residues). Peptides III and IV were identified by mass, with the measured value compared with predicted value.

sequence resulted in lower production and elevated degradation (fragmentation) of the protein (data not shown). The partial DnaE-n sequence is termed DnaE-n' and consisted of a portion of the Ext-n sequence (216 aa, proximal to the intein) followed by the entire Int-n sequence. The DnaE-c- and DnaE-n'-coding sequences were separated by a small intergenic spacer that contained a Shine-Dalgarno sequence (ribosome-binding site) followed by an AT-rich sequence. The DnaE-c-coding sequence was placed in front of the DnaE-n-coding sequence, preventing accidental fusion of the split intein sequences, which might arise through accidental translation of the small intergenic spacer.

E. coli cells containing the above recombinant plasmid were induced to produce the DnaE-c protein, the DnaE-n' protein, and possibly a spliced protein. Three protein products (C, N, and N-C) were observed after the induction (Fig. 3*B*). Protein C and protein N were identified as the precursor proteins DnaE-c and DnaE-n', respectively. Their apparent sizes matched closely the predicted sizes (51 kDa for C and 38 kDa for N), and each of them was recognized specifically by antiserum raised against that protein. The third protein, N-C, was identified as a spliced protein (ligated exteins). First, its apparent size matched closely the predicted size of a spliced protein (71 kDa). Second, protein N-C was recognized by both the anti-N and the anti-C antisera, indicating that it contains both DnaE-n and DnaE-c sequences. Finally, protein N-C was firmly identified as the spliced protein by protein sequencing and peptide analysis (Fig. 3*C*). N-terminal protein sequencing of protein N-C revealed a 17-aa sequence, KMDFLGLKNLTTLQRAV, which matched precisely the predicted DnaE-n' sequence at amino acid positions 5–21. Amino acids at positions 2–4 were not determined, because of sequencing failures at these positions, and the N-terminal f-Met apparently had been removed in the *E. coli* cell. The protein N-C was further treated with protease trypsin, and the resulting polypeptides were selectively analyzed. Two polypeptides (peptides III and IV) inside the DnaE-c sequence were identified by matching their molecular masses to predicted molecular masses. Peptide III corresponded to the sequence SHSTAYAYVTYQTAYLK (amino acid positions 220–236), whereas peptide IV corresponded to the sequence EHLGFYVSEHPLK (amino acid positions 428–440). Most importantly, a polypeptide (peptide II) spanning the spliced junction was identified and sequenced. Its sequence, FAEYCFNK, matches precisely the predicted sequence in a spliced protein, with the sequence FAEY being the last four residues of Ext-n' and the sequence CFNK being the first four residues of Ext-c. This shows precise excision of the intein sequences (Int-n and Int-c) and joining of the extein sequences (Ext-n' and Ext-c) by a normal peptide bond. The two excised intein fragments were predicted but not observed, most likely because of their small sizes (14 kDa for Int-n and 4 kDa for Int-c), weak binding by the anti-N and anti-C antisera, and/or rapid degradation in the *E. coli* cell. Nevertheless, production of the spliced protein (protein N-C) demonstrates that protein *trans*-splicing had occurred. Comparing the amount of protein N-C and the amount of protein N indicates that ≈80% of the precursor protein N was incorporated into the spliced protein. The remaining protein N may have misfolded. Protein C accumulated much more than protein N, indicating that the *dnaE-c* gene was expressed much more than the downstream *dnaE-n'* gene. This may be because of inefficient translational coupling of the two-gene operon or a more rapid degradation of protein N.

DISCUSSION

The *Ssp* DnaE intein is identified as a naturally occurring split mini-intein in *Synechocystis sp.* PCC6803, and it is shown to be capable of protein *trans*-splicing. The two DnaE-like genes, *dnaE-n* and *dnaE-c*, are clearly two members of an intein-containing split DnaE gene, with the split being inside the intein-coding sequence. Protein sequences deduced from the split DnaE gene, after excluding the intein sequences, reconstitute a complete DnaE protein that has neither gap nor overlapping sequences at the split point. It also has the expected degrees of sequence identity to the continuous DnaE sequences of other bacterial organisms. The two intein sequences, Int-n and Int-c, not only have intein-like sequence features but also are proven to be two parts of a split intein by demonstrating a protein *trans*-splicing activity in *E. coli* cells. This *Ssp* DnaE intein, consisting of two separate polypeptides with a composite size of 159 aa, represents a split mini-intein that is apparently capable of forming a functional splicing domain. Four conserved sequence blocks (A, B, F, G) have been previously localized in the splicing domain of inteins (6, 10, 15, 22, 27, 37). All of the four sequence blocks appear to exist in the *Ssp* DnaE intein (Fig. 2*B*), with blocks A and B located on Int-n, with blocks F and G located on Int-c. The *Ssp* DnaE intein lacks a highly conserved His residue (replaced by Ala) immediately before the C-terminal Asn. Four other inteins (*Ceu* ClpP, *Mja* PEP, *Mja* KlbA, and *Mja* RpolA′) also lack this penultimate His, in which the His is replaced by Gly, Ser, or Phe. This His has been shown to assist in Asn cyclization leading to cleavage of the peptide bond between intein and C-extein (17), and efficient splicing of the *Ceu* ClpP intein in *E. coli* cells required a restoration of this His residue (35). The observation of *trans*-splicing activity with the *Ssp* DnaE intein shows that this His residue is not required for protein splicing of this intein.

The finding of a split mini-intein has implications on intein evolution. The *Ssp* DnaE intein likely evolved from a continuous intein that later lost its sequence continuity. This result probably occurred through one or more genomic rearrangement events that separated the two halves of the DnaE gene (*dnaE-n* and *dnaE-c*) to different parts of the genome. A possible progenitor DnaE intein has not been found, and the 30% sequence identity between *Ssp* DnaE intein and the *Rma* DnaB intein (present in a DNA helicase) may be a coincidence, considering that the two inteins have nonhomologous exteins and dissimilar insertion sites. Emergence of a split intein requires that it possesses protein *trans*-splicing activity, unless the exteins can function without ligation and without removing the intein sequences. Other inteins also may possess a potential of becoming split inteins, as protein *trans*-splicing has been demonstrated with intein fragments engineered from several continuous inteins (36, 37, 40, 41). The *Ssp* DnaE intein (in fragments) has a total size of a mini-intein (splicing domain only) and lacks any of the endonuclease sequence motifs. The *Ssp* DnaE intein, like other inteins that lack an endonuclease domain, may once have had and lost the endonuclease domain (13), or alternatively it may never have acquired an endonuclease domain. The split site in the *Ssp* DnaE intein coincides with predicted endonuclease insertion site, indicating that this site of the intein is tolerant of both insertion and cleavage. If the *Ssp* DnaE intein once had and lost its endonuclease domain, this could have occurred before or after the loss of sequence continuity. An intein presumably loses the ability of intein homing once the endonuclease domain is lost. As for the *Ssp* DnaE intein, having the two intein fragments on different parts of the genome would prevent intein homing even if the endonuclease domain were present.

The *Ssp* DnaE intein likely does protein *trans*-splicing in its native cyanobacterial cell, as it did so in *E. coli* cells. A DnaE protein, either a spliced protein or precursors, has not been detected in the total protein of *Synechocystis* sp. PCC6803 by using the available anti-DnaE antisera (data not shown). This is most likely because of a combination of weak antisera and low levels of the DnaE protein. DnaE has been known to exist at very low levels in other bacterial cells. The *E. coli* DnaE protein was estimated at 10–12 molecules per cell (38), which is sufficient to replicate the *E. coli* genome approximately every 0.5 hr. In comparison, *Synechocystis* sp. PCC6803 has a smaller genome that needs to be duplicated only every 10 hr (approximate cell-doubling time). It is therefore not unreasonable for this organism to have extremely low levels of the DnaE protein for DNA replication. Nevertheless, a DnaE protein is essential for the cell, and there is no other DnaE-like gene (complete or partial) beside *dnaE-n* and *dnaE-c* in this genome. These two genes, unlike pseudo genes, maintain long ORFs (2,694 bp for *dnaE-n* and 1,377 bp for *dnaE-c*), whereas their noncoding frames have numerous termination codons. Production of a functional DnaE protein likely requires protein *trans*-splicing to remove the intein sequences and ligate the extein sequences. It is less likely, although possible, for the two precursor proteins (DnaE-n and DnaE-c) to reconstitute a functional protein without splicing, considering that the intein sequences interrupt both the β-binding domain and the τ-binding domain. Although the polymerase active site is contained within the DnaE-n precursor protein, both the β-binding domain and the τ-binding domain are interrupted by the intein sequences and split between the DnaE-n and DnaE-c precursor proteins. There is no indication that the half intein sequences (Int-n and Int-c) can be cleaved off the precursor proteins without undergoing protein *trans*-splicing. Such a cleavage product was not observed with the DnaE-n and DnaE-c proteins in *E. coli*. Half inteins engineered *in vitro* from other inteins also lack such a cleavage activity (36, 37). Functional β- and τ-binding domains are essential, because interactions of DnaE with the β subunit (DNA clamp) and the τ subunit are critical for the function of DNA polymerase III (30).

Protein *trans*-splicing has been demonstrated with engineered inteins *in vivo* and *in vitro* (36, 37, 40, 41) and has produced insights into the structural requirements for protein splicing. The discovery of the *Ssp* DnaE intein, a natural split intein that does protein *trans*-splicing, provides a new perspective on this phenomenon. In terms of structural requirements for protein splicing, the size and sequence of this naturally evolved split mini-intein are in close agreement with those of the smallest functional mini-inteins that have been engineered so far in a laboratory (15, 41). In terms of possible biological function, the *trans*-splicing reaction between the DnaE-n and DnaE-c precursor proteins may present a step in which the synthesis of a functional DnaE protein is regulated. Absence of the penultimate C-terminal His residue (replaced by Ala) in the *Ssp* DnaE intein, although not preventing protein *trans*-splicing, may slow down the splicing reaction, as was the case for other inteins (16, 17, 35). A slow and regulated splicing step may be a mechanism for assuring very low levels of production of the mature DnaE protein. The β and τ subunits of DNA polymerase III bind strongly with the DnaE protein and may therefore affect the *trans*-splicing reaction by bringing together the two precursor polypeptides of DnaE. It is interesting that the τ subunit of this organism also has an intein (*Ssp* DnaX intein), although the *Ssp* DnaX intein has a continuous sequence and is not specifically related to the *Ssp* DnaE intein in sequence and insertion site (29).

This work was supported by a grant from the Medical Research Council of Canada.

1. Perler, F. B., Davis, E. O., Dean, G. E., Gimble, F. S., Jack, W. E., Neff, N., Noren, C. J., Thorner, J. & Belfort, M. (1994) *Nucleic Acids Res.* **22**, 1125–1127.

2. Perler, F. B. (1998) *Cell* **92,** 1–4.
3. Kane, P. M., Yamashiro, C. T., Wolczyk, D. F., Neff, N., Goebl, M. & Stevens, T. H. (1990) *Science* **250,** 651–657.
4. Colston, M. J. & Davis, E. O. (1994) *Mol. Microbiol.* **12,** 359–363.
5. Cooper, A. A. & Stevens, T. H. (1995) *Trends Biochem. Sci.* **20,** 351–356.
6. Perler, F. B., Olsen, G. J. & Adam, E. (1997) *Nucleic Acids Res.* **25,** 1087–1093.
7. Dalgaard, J. Z., Moser, M. J., Hughey, R. & Mian, I. S. (1997) *J. Comp. Biol.* **4,** 193–214.
8. Dalgaard, J. Z., Klar, A. J., Moser, M. J., Holley, W. R., Chatterjee, A. & Mian, I. S. (1997) *Nucleic Acids Res.* **25,** 4626–4638.
9. Pietrokovski, S. (1998) *Protein Sci.* **7,** 64–71.
10. Duan, X., Gimble, F. S. & Quiocho, F. A. (1997) *Cell* **89,** 555–564.
11. Kawasaki, M., Nogami, S., Satow, Y., Ohya, Y. & Anraku, Y. (1997) *J. Biol. Chem.* **272,** 15668–15674.
12. Nogami, S., Satow, Y., Ohya, Y. & Anraku, Y. (1997) *Genetics* **147,** 73–85.
13. Telenti, A., Southworth, M., Alcaide, F., Daugelat, S., Jacobs, W. R. Jr. & Perler, F. B. (1997) *J. Bacteriol.* **179,** 6378–6382.
14. Chong, S. & Xu, M.-Q. (1997) *J. Biol. Chem.* **272,** 15587–15590.
15. Derbyshire, V., Wood, D. W., Wu, W., Dansereau, J. T., Dalgaard, J. Z. & Belfort, M. (1997) *Proc. Natl. Acad. Sci. USA* **94,** 11466–11471.
16. Shao, Y., Xu, M. Q. & Paulus, H. (1996) *Biochemistry* **35,** 3810–3815.
17. Xu, M.-Q. & Perler, F. B. (1996) *EMBO J.* **15,** 5146–5153.
18. Chong, S., Shao, Y., Paulus, H., Benner, J. & Perler, F. B. (1996) *J. Biol. Chem.* **271.**
19. Xu, M. Q., Southworth, M. W., Mersha, F. B., Hornstra, L. J. & Perler, F. B. (1993) *Cell* **75,** 1371–1377.
20. Xu, M. Q., Comb, D. G., Paulus, H., Noren, C. J., Shao, Y. & Perler, F. B. (1994) *EMBO J.* **13,** 5517–5522.
21. Shao, Y., Xu, M. Q. & Paulus, H. (1995) *Biochemistry* **34,** 10844–10850.
22. Klabunde, T., Sharma, S., Telenti, A., Jacobs Jr., W. R. & Sacchettini, J. C. (1998) *Nat. Struct. Biol.* **5,** 31–36.
23. Lambowitz, A. M. & Belfort, M. (1993) *Annu. Rev. Biochem.* **62,** 587–622.
24. Gimble, F. S. & Thorner, J. (1992) *Nature (London)* **357,** 301–306.
25. Shub, D. A. & Goodrich-Blair, H. (1992) *Cell* **71,** 183–186.
26. Doolittle, R. F. (1993) *Proc. Natl. Acad. Sci. USA* **90,** 5379–5381.
27. Pietrokovski, S. (1994) *Protein. Sci.* **3,** 2340–2350.
28. Pietrokovski, S. (1996) *Trends Genet.* **12,** 287–288.
29. Liu, X.-Q. & Hu, Z. (1997) *FEBS Lett.* **408,** 311–314.
30. Kim, D. R. & McHenry, C. S. (1996) *J. Biol. Chem.* **271,** 20699–20704.
31. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990) *J. Mol. Biol.* **215,** 403–410.
32. Thompson, J. D., Higgins, D. G. & Gibson, T. J. (1994) *Nucleic Acids Res.* **22,** 4673–4680.
33. Kaneko, T., Sato, S., Kotani, H., Tanaka, A., Asamizu, E., Nakamura, Y., Miyajima, N., Hirosawa, M., Sugiura, M., Sasamoto, S., *et al.* (1996) *DNA Res.* **3,** 109–136.
34. Liu, X.-Q. & Hu, Z. (1997) *Proc. Natl. Acad. Sci. USA* **94,** 7851–7856.
35. Wang, S. & Liu, X.-Q. (1997) *J. Biol. Chem.* **272,** 11869–11873.
36. Southworth, M. W., Adam, E., Panne, D., Byer, R., Kautz, R. & Perler, F. B. (1998) *EMBO J.* **17,** 918–926.
37. Shingledecker, K., Jiang, S.-Q. & Paulus, H. (1998) *Gene* **207,** 187–195.
38. Wu, Y. H., Franden, M. A., Hawker, J. R. & McHenry, C. S. (1984) *J. Biol. Chem.* **259,** 12117–12122.
39. Gorbalenya, A. E. (1998) *Nucleic Acids Res.* **26,** 1741–1748.
40. Mills, K. V., Lew, B. M., Jiang, S.-Q. & Paulus, H. (1998) *Proc. Natl. Acad. Sci. USA* **95,** 3543–3548.
41. Wu, H., Xu, M.-Q. & Liu X.-Q. (1998) *Biochim. Biophys. Acta*, in press.

Exhibit F

The EMBO Journal Vol.17 No.4 pp.918–926, 1998

Control of protein splicing by intein fragment reassembly

Maurice W.Southworth, Eric Adam, Daniel Panne[1], Robyn Byer, Roger Kautz[2] and Francine B.Perler[3]

New England Biolabs, 32 Tozer Road, Beverly, MA 01915, USA

[1]Present address: Biozentrum, Abteilung Mikrobiologie, Klingelbergstrasse 70, CH-4056 Basel, Switzerland
[2]Present address: Barnett Institute, Northeastern University, Boston, MA 02115, USA

[3]Corresponding author
e-mail: perler@neb.com

Inteins are protein splicing elements that mediate their excision from precursor proteins and the joining of the flanking protein sequences (exteins). In this study, protein splicing was controlled by splitting precursor proteins within the Psp Pol-1 intein and expressing the resultant fragments in separate hosts. Reconstitution of an active intein was achieved by *in vitro* assembly of precursor fragments. Both splicing and intein endonuclease activity were restored. Complementary fragments from two of the three fragmentation positions tested were able to splice *in vitro*. Fragments resulting in redundant overlaps of intein sequences or containing affinity tags at the fragmentation sites were able to splice. Fragment pairs resulting in a gap in the intein sequence failed to splice or cleave. However, similar deletions in unfragmented precursors also failed to splice or cleave. Single splice junction cleavage was not observed with single fragments. *In vitro* splicing of intein fragments under native conditions was achieved using mini exteins. *Trans*-splicing allows differential modification of defined regions of a protein prior to extein ligation, generating partially labeled proteins for NMR analysis or enabling the study of the effects of any type of protein modification on a limited region of a protein.

Keywords: intein/protein expression/reconstitution/split proteins/thermophile/urea

Introduction

Protein splicing is a post-translational process that results in excision of an intein (protein splicing element) from a precursor protein and the ligation of the flanking protein sequences (exteins) to yield two mature proteins, the intein and the ligated exteins (Perler *et al.*, 1994). The native peptide bond formed between the exteins (Cooper *et al.*, 1993) distinguishes protein splicing from other forms of autoprocessing (Perler *et al.*, 1997b). The self-catalytic reaction requires four nucleophilic displacements mediated by three conserved splice junction residues: (i) a Ser or Cys at the intein N-terminus; (ii) an Asn at the intein C-terminus; and (iii) a Ser, Thr or Cys at the beginning of the C-extein (Xu *et al.*, 1993, 1994; Shao *et al.*, 1995, 1996; Chong *et al.*, 1996; Xu and Perler, 1996). Genetic, biochemical and structural studies have shown that formation of the splicing active site requires proper folding of the intein to bring together the two splice junctions that can be >500 amino acids apart, plus other intein residues that may assist in the nucleophilic displacements, such as the conserved His in intein blocks B and G (Pietrokovski, 1994, 1996; Duan *et al.*, 1997; Hall *et al.*, 1997; Kawasaki *et al.*, 1997; Perler *et al.*, 1997a,b).

Many inteins are bifunctional proteins, having both splicing and homing endonuclease activity (Bremer *et al.*, 1992; Perler *et al.*, 1992; Mueller *et al.*, 1994). The mature Psp Pol-1 intein is also a homing endonuclease, PI-PspI, that specifically cleaves the intein insertion site in DNA polymerase genes that lack the intein (F.Perler, unpublished data). This type of homing endonuclease activity is thought to initiate intein gene mobility into inteinless extein alleles (Mueller *et al.*, 1994; Perler *et al.*, 1997a).

In an attempt to control splicing and allow differential labeling or modification of portions of a protein, we split several precursors within the Psp Pol-1 intein and examined whether splicing could be reconstituted *in vitro* from the separately purified parts (Figure 1). Limited proteolysis experiments have proven that folded proteins can remain active despite the presence of breaks in the peptide backbone (Anfinsen and Scheraga, 1975). Previous studies have also indicated that under certain conditions, protein fragments are able to find their complementary partners and fold properly to generate an active enzyme (Kato and Anfinsen, 1969; Matsuyama *et al.*, 1990; Burbaum and Schimmel, 1991; Sancho and Fersht, 1992; Kanaya and Kanaya, 1995; Tasayco and Chao, 1995; Gross *et al.*, 1996). These and other studies also demonstrated that the conformation of protein fragments is often disordered, and hydrophobic regions that are normally buried in the intact protein may be exposed, leading to aggregation, insolubility or *in vivo* proteolysis. Most *in vitro* assembly protocols include a denaturation step prior to or during fragment association (Kato and Anfinsen, 1969; Anfinsen and Scheraga, 1975; Matsuyama *et al.*, 1990; Burbaum and Schimmel, 1991; Sancho and Fersht, 1992; Kanaya and Kanaya, 1995; Tasayco and Chao, 1995). Finally, *in vivo* reconstitution is often more efficient than *in vitro* reconstitution (Gross *et al.*, 1996). *In vivo* reassembly can be aided by reassociation before the co-expressed fragments misfold and/or by the assistance of the powerful protein folding machinery present in the cell. However, *in vivo* assembly does not allow differential labeling or modification of portions of a protein, nor does it necessarily block protein splicing *in vivo*.

In this study, we examined the ability of an enzyme from an extreme thermophile (*Pyrococcus* species, isolate

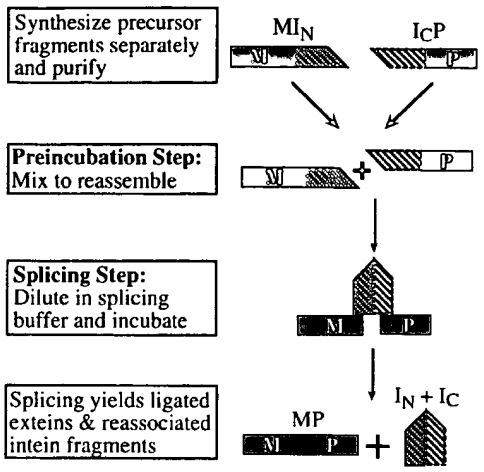


Fig. 1. Scheme for reassembly of split inteins. The MIP protein splicing precursor (Xu *et al.*, 1993) was split at various locations within the intein (I) to yield N-terminal fragments (MI_N) and C-terminal fragments (I_CP). However, precursors with other N-exteins or C-exteins can likewise be split within the intein. Purification tags can also be added to the intein split spites. After purification, the two fragments are mixed in buffers containing various amounts of urea (0–8.0 M) and allowed to reassemble. The reconstituted intein then directs the splicing reaction, resulting in joining of the extein fragments with a native peptide bond. The reconstituted intein fragments also display PI-PspI endonuclease activity.

GB-D) to assemble at temperatures of up to 100°C below its normal synthesis and folding temperatures. Assembly into an active enzyme was monitored by assaying protein splicing or endonuclease activity. The ability to reassemble intein fragments into an active enzyme converts any intein into a controllable protein splicing element. Moreover, it provides a method for the specific labeling or modification (e.g. phosphorylation, glycosylation, acetylation) of a protein fragment prior to assembly, allowing the analysis of the effects of the specific modification.

Results

Construction and expression of split precursor proteins

Intein fragmentation studies were performed with the previously characterized chimeric protein MIP which is a three part fusion of the *Escherichia coli* maltose-binding protein (M or MBP, the N-extein), the Psp Pol-1 intein (I) and the ΔSal fragment of *Dirofilaria immitis* paramyosin (P, the C-extein) (Xu *et al.*, 1993). Splicing of MIP is optimal at pH 6–7 with a half-time of 20–30 min *in vitro* and is inhibited at low temperatures (4–16°C) or pH values above pH 9 (Xu *et al.*, 1993, 1994; Xu and Perler, 1996).

Since there are no precise rules for choosing split sites (the position at which the protein is split into fragments) (Matsuyama *et al.*, 1990; Burbaum and Schimmel, 1991; Gross *et al.*, 1996), three positions within the Psp Pol-1 intein were tested. No structural information was available for any intein or homing endonuclease at the inception of this project, although sequences of several alleles of the Psp pol-1 intein were available for comparison (Perler *et al.*, 1997a). We hypothesized that non-conserved, unstructured surface locations might be less essential to the intein and, therefore, breakage of the peptide backbone in these regions might be less detrimental. Therefore, the three split sites were chosen in highly variable regions of the Psp Pol-1 intein that were predicted by computer modeling (Rost *et al.*, 1994) to be in unstructured loops with potential surface locations. Precursor proteins were split following intein residues Glu108 (MI_{N1} and $I_{C1}P$), Leu249 (MI_{N2} and $I_{C2}P$) and Arg440 (MI_{N3} and $I_{C3}P$) at split sites 1, 2 and 3, respectively (Figure 2). Leu249 precedes a naturally occurring Met at position 250, and Arg440 is near a protease-sensitive site at Lys442 (J.Benner and T.Davis, personal communication).

N-terminal precursor fragments (MI_{N1}, MI_{N2} and MI_{N3}) were synthesized as soluble proteins (10–40 mg/l) while C-terminal precursor fragments ($I_{C2}P$, $I_{C2}P\Delta$ and $I_{C3}P$) were synthesized as insoluble proteins (30–40 mg/l). $I_{C1}P$ was very sensitive to *in vivo* proteolysis, but small amounts of full-length $I_{C1}P$ protein (0.5 mg/l) could be isolated under certain induction conditions. To eliminate the possibility that the insolubility of I_CP fragments was due to the paramyosin domain, the paramyosin extein was replaced in $I_{C3}P$ by *E.coli* thioredoxin or the chitin-binding domain from the *Bacillus circulans* chitinase; both new C-terminal fragments were also insoluble.

Time course of splicing in trans confirms the protein splicing pathway

Splicing of precursor fragments *in trans* was successful when MIP was split at sites 2 and 3, after Leu249 or Arg440, respectively, but not at site 1, after Glu108 (Figures 2–5 and Table I). This percentage of successful fragment reassembly is similar to that reported in other systems (Matsuyama *et al.*, 1990; Burbaum and Schimmel, 1991). Products that would be unique to cleavage reactions were not observed. For example, I_N is a potential product of both splicing and cleavage while free M can only result from cleavage of MI_N. Similar results were also observed with intact MIP, where cleavage only occurred *in vivo* or after mutagenesis and not *in vitro*.

Splicing *in trans* confirmed the order of splice junction cleavage in the protein splicing pathway since such cleavage releases identifiable intein fragments in SDS–PAGE. The protein splicing pathway begins with ester formation followed by cleavage at the N-terminal splice site resulting in formation of a slowly migrating branched intermediate (MIP*) containing M connected to the side chain of S538 in IP (Xu *et al.*, 1993, 1994; Shao *et al.*, 1995, 1996; Xu and Perler, 1996). In *trans*-splicing reactions, N-terminal splice site cleavage would be detected in SDS–PAGE by the appearance of I_N and the branched intermediate, MI_CP^*. The next step in the pathway is resolution of the branched intermediate by Asn cyclization which would lead to cleavage at the C-terminal splice site. In *trans*-splicing, this would result in the release of I_C and the ligated exteins (MP) from MI_CP^*. The MI_{N3} plus $I_{C3}P$ time course shown in the left panel of Figure 3 illustrates this splicing pathway. The first products observed are the branched intermediate ($MI_{C3}P^*$) and cleaved I_{N3}. Small amounts of both were present after the overnight incubation in 3.6 M urea (0 min splicing reaction sample). As the splicing reaction continued, more branched intermediate and I_{N3} were formed, followed by the appearance of the spliced product (MP) and I_{C3}. By 90 min, most of $MI_{C3}P^*$

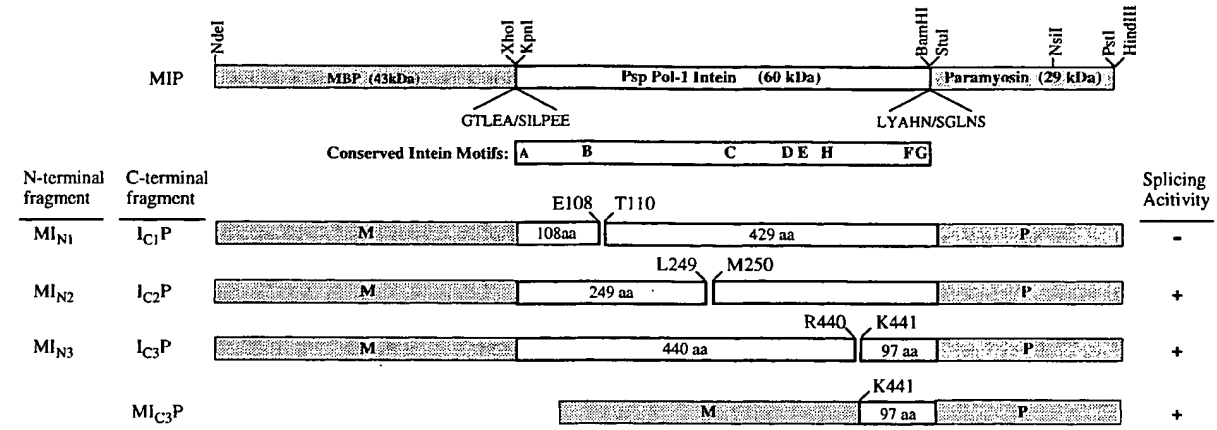


Fig. 2. Map of MIP fragments. Restriction enzyme sites within the MIP gene used to construct subclones are shown across the top of the MIP precursor, and residues surrounding the Psp Pol-1 intein splice junctions (/) are shown below the MIP precursor. The eight conserved intein motifs are depicted, including splicing motifs (blocks A, B, F and G) and endonuclease motifs (blocks C, D, E and H). MIP was split at three sites after intein amino acids E108, L249 and R440 to generate three sets of complementary fragments. Fragment names are listed to the left of each fragment pair and include an N or C subscript indicating an N- or C-terminal intein sequence, respectively, and a subscript split site number (1–3). The terminal intein residue at each split site is shown above the fragment, and the number of intein amino acids in each fragment is listed within the white intein box. *In vitro* splicing activity of complementary fragments is shown to the right of each fragment pair. An MBP affinity tag was fused to the N-terminus of $I_{C3}P$ to generate $MI_{C3}P$. Splicing activity of $MI_{C3}P$ was assayed with its complementary fragment, MI_{N3}. Abbreviations: M or MBP, maltose-binding protein; P, paramyosin; (/) splice junction; white box, intein or intein fragment; shaded box, MBP or paramyosin exteins.

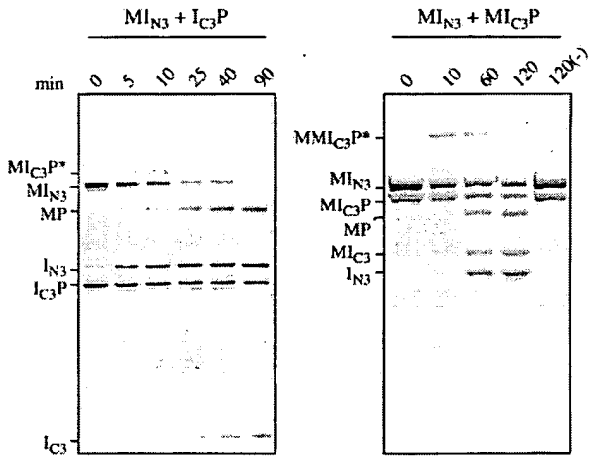


Fig. 3. Time course of splicing with MIP split at site 3. Splicing of the MI_{N3} plus $I_{C3}P$ complementary pair of fragments proceeded as predicted by the protein splicing mechanism, including formation of a slowly migrating intermediate ($MI_{C3}P^*$, left panel or $MMI_{C3}P^*$, right panel), irrespective of the presence of an MBP affinity tag at the split site in $MI_{C3}P$. Left panel: MI_{N3} was mixed with $I_{C3}P$ in 3.6 M urea buffer at 4°C overnight (protocol 2) and then diluted 10-fold into splicing buffer followed by incubation at 37°C for 0–90 min. Right panel: MI_{N3} was mixed with $MI_{C3}P$ containing an MBP tag at the split site, treated as above and incubated for 0–120 min at 37°C. Lane 120(–) was not pre-incubated, but instead the fragments in amylose column elution buffer were diluted directly into splicing buffer and incubated at 37°C immediately after fragment mixing. Abbreviations: MI_{N3} and $I_{C3}P$ or $MI_{C3}P$, substrate fragments; MP, spliced product; I_{N3}, I_{C3} or MI_{C3}, intein fragment products. The SDS–PAGE gels were stained with Coomassie blue.

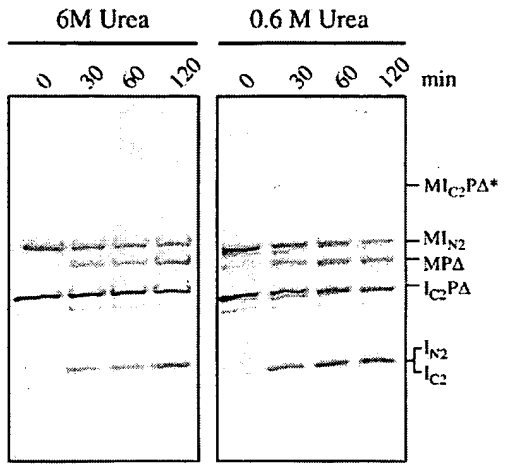


Fig. 4. Effect of urea in the splicing buffer. MI_{N2} and $I_{C2}P\Delta$ complementary fragments were pre-incubated in 6.0 M urea buffer at 4°C and then diluted 10-fold into splicing buffer containing 6 M urea (protocol 3, left panel) or 0 M urea (protocol 1, right panel) followed by incubation at 37°C for 0–120 min. The SDS–PAGE gel was stained with Coomassie blue. The presence of urea in the splicing buffer had no significant effect on the production of spliced MPΔ.

and the MI_{N3} substrate had disappeared, leaving some unreacted $I_{C3}P$ substrate which was in molar excess.

Conditions for functional assembly of split precursors

Most *in vitro* protocols for assembly of split proteins involve denaturation followed by renaturation (Kato and Anfinsen, 1969; Matsuyama *et al.*, 1990; Burbaum and Schimmel, 1991; Sancho and Fersht, 1992; Kanaya and Kanaya, 1995; Tasayco and Chao, 1995; Gross *et al.*, 1996). Several protocols for intein fragment assembly were examined (Table I). Parameters that were varied in this study included urea concentration (0–8.0 M), temperature (4 or 37°C) and length (0–20 h) of the pre-incubation step, and urea concentration (0–8.0 M) of the splicing step. The splicing step consisted of renaturation by rapidly diluting the pre-incubation mixtures ≥10-fold in splicing buffer (without urea, except in protocol 3) and incubating at 37°C to stimulate splicing. Splicing was monitored by observing the disappearance of MI_N and I_CP substrates and the appearance of MP, I_N and I_C products.

However, splicing efficiency was calculated only with respect to the synthesis of MP, since this is the desired spliced product. Significant amounts of either or both substrate fragments often remained at the end of the splicing reaction, partially because an equal amount of each substrate does not represent equal moles of fragments (due to differences in molecular weight). Misfolding and/or aggregation of substrate may also contribute to the failure of splicing reactions to go to 100%, since a 4-fold molar excess of one fragment could not drive the splicing of the second fragment to completion (data not shown). This result suggests that some fraction of each fragment is incapable of splicing.

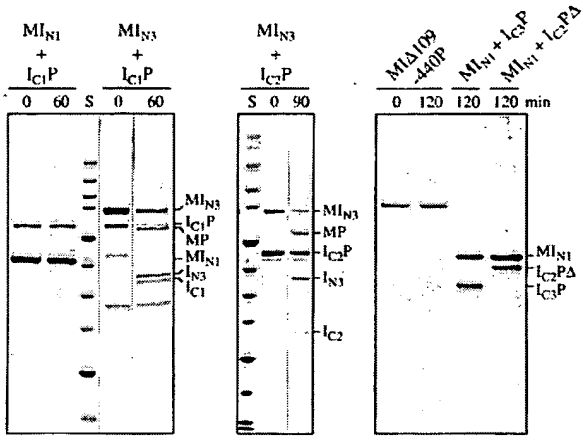


Fig. 5. Splicing with fragment pairs resulting in gaps or overlaps and with MIP intein deletions. Left panel: splicing of $I_{C1}P$ with MI_{N1} or MI_{N3} using protocol 2 after 0 or 60 min incubations in splicing buffer. Splicing of MI_{N3} plus $I_{C1}P$ (330 amino acid overlap) results in production of MP, I_{N3} and I_{C1}. Middle panel: splicing of MI_{N3} plus $I_{C2}P$ containing an overlap of 190 amino acids. Right panel: fragment pairs resulting in a gap in the intein sequence (MI_{N1} plus either $I_{C3}P$ or $I_{C2}P\Delta$) and a MIP precursor with a deletion of intein residues 109–440 (MIΔ109–440P) which mimics the gap in the MI_{N1} plus $I_{C3}P$ pair failed to splice or cleave after protocol 2 and incubation in splicing buffer for 120 min. Lane S, molecular weight standards.

Five standard protocols were employed that consisted of (i) having one, both or neither fragment in urea prior to the pre-incubation step, (ii) a pre-incubation step at different urea concentrations, and (iii) a splicing step at different urea concentrations (Table I). The initial urea concentration of the separate fragments had no effect on splicing, but the urea concentration in the pre-incubation buffer drastically affected splicing. There was little difference in MP formation when there was 3.0–7.2 M urea in the pre-incubation buffer, but splicing was blocked or inhibited in pre-incubation buffers containing 0–1.8 M urea. No splicing was observed if the pre-incubation mix was diluted immediately into splicing buffer. Splicing efficiency improved with increasing pre-incubation times up to 4 h, after which there was only a small increase in spliced product. Allowing the diluted pre-incubated samples to 'renature' in splicing buffer at 4°C for 0–12 h before shifting to 37°C had no effect on splicing efficiency (data not shown). The presence of ≤6.0 M urea in the splicing buffer had little effect on splicing efficiency (Figure 4), but splicing was blocked in 8.0 M urea after 4 h at 37°C (Table I). Use of $I_{C3}P$ or MI_{N3} crude extracts had no significant effect on splicing, indicating that reassembly could occur in the presence of exogenous proteins (data not shown). Varying the pH (5.5 or 7.9) of the pre-incubation and splicing buffers had no effect on splicing of MI_{N3} plus $I_{C3}P$ (data not shown). The effects of several folding aids were tested in the pre-incubation or splicing buffers, or both. Triton X-100 (1%), glycerol (10%), PEG 8000 (0.3%), arginine (0.5 M) and SDS

Table I. Comparison of pre-incubation and splicing reaction conditions with complementary fragment pairs

Protocol[a]	MI fragment (M urea)	IP fragment (M urea)	Pre-incubation M urea	Pre-incubation time	Splicing reaction M urea	Splicing efficiency[b]
Splicing efficiency versus splicing protocol						
1	MI_{N2} (6.0 M)	$I_{C2}P$ (6.0 M)	6.0 M	4 h	0.6 M	47 ± 3%
2	MI_{N2} (0 M)	$I_{C2}P$ (7.2 M)	3.6 M	4 h	0.36 M	53 ± 9%
2	MI_{N3} (0 M)	$MI_{C3}P$ (0 M)	3.6 M	4 h	0.36 M	54 ± 6%
3	MI_{N2} (6.0 M)	$I_{C2}P$ (6.0 M)	6.0 M	4 h	6.0 M	52 ± 3%
3	MI_{N3} (0 M)	$MI_{C3}P$ (0 M)	8.0 M	4 h	8.0 M	0%
4	MI_{N3} (0 M)	$MI_{C3}P$ (0 M)	0 M	4 h	0 M	0%
Splicing efficiency versus complementary fragment pair						
2	MI_{N1} (0 M)	$I_{C1}P$ (7.2 M)	3.6 M	4 h	0.36 M	0%
2	MI_{N2} (0 M)	$I_{C2}P$ (7.2 M)	3.6 M	4 h	0.36 M	53 ± 9%
2	MI_{N3} (0 M)	$I_{C3}P$ (7.2 M)	3.6 M	4 h	0.36 M	59 ± 6%
2	MI_{N3} (0 M)	$MI_{C3}P$ (0 M)	3.6 M	4 h	0.36 M	54 ± 6%
Splicing efficiency versus time of pre-incubation						
2	MI_{N2} (0 M)	$I_{C2}P$ (7.2 M)	3.6 M	0 h	0.36 M	0%
2	MI_{N2} (0 M)	$I_{C2}P$ (7.2 M)	3.6 M	0.5 h	0.36 M	23 ± 6%
2	MI_{N2} (0 M)	$I_{C2}P$ (7.2 M)	3.6 M	4 h	0.36 M	53 ± 9%
2	MI_{N2} (0 M)	$I_{C2}P$ (7.2 M)	3.6 M	20 h	0.36 M	74 ± 5%
Splicing efficiency versus urea concentration in the pre-incubation reaction						
4	MI_{N3} (0 M)	$MI_{C3}P$ (0 M)	0 M	4 h	0 M	0%
5	MI_{N3} (0 M)	$I_{C3}P$ (7.2 M)	0.9 M	4 h	0.09 M	0%
5	MI_{N3} (0 M)	$I_{C3}P$ (7.2 M)	1.8 M	4 h	0.18 M	9 ± 1%
2	MI_{N3} (0 M)	$I_{C3}P$ (7.2 M)	3.6 M	4 h	0.36 M	59 ± 6%

[a]Fragments were mixed and pre-incubated at 4°C for the indicated times and urea concentrations. Pre-incubated samples were then diluted into splicing buffer at the indicated final urea concentrations and immediately incubated at 37°C for 2 h.
[b]Splicing efficiency = (moles of MP produced/initial moles of limiting substrate)×100 and was calculated from two or more independent experiments.

(0.01%) had no effect on the rate of splicing or the amount of spliced product observed after a 2 h splicing reaction at 37°C. However, 0.1% SDS in the pre-incubation buffer blocked splicing.

Functional assembly of split inteins reconstitutes PI-PspI endonuclease activity

The MI_{N2} plus $I_{C2}P$ pair was also tested for endonuclease activity. After a 2 h splicing reaction including pre-treatment in urea (protocol 1), the reassembled and spliced MI_{N2} plus $I_{C2}P$ sample was added to a standard PI-PspI digestion mixture. The reconstituted intein yielded the same cleavage pattern as the MIP52 control, although in both cases the amount of enzyme added was insufficient to yield a complete digest (Figure 6, lanes 2 and 3). Cleavage was dependent on pre-incubation in urea, since no digestion was observed if the fragment pair was added directly to the DNA digestion mixture (Figure 6, lane 1). These results suggest that the presence of the DNA substrate does not obviate the need for the urea pre-incubation step in fragment assembly. The individual fragments did not have detectable endonuclease activity (Figure 6, lanes 4 and 5) despite the fact that the $I_{C2}P$ fragment begins 30 amino acids N-terminal to intein block C and contains all of the putative homing endonuclease motifs (intein blocks C, D, E and H) (Mueller *et al.*, 1994; Pietrokovski, 1994; Perler *et al.*, 1997a). With the information presently available, it is difficult to correlate the structure of the Sce VMA intein with amino acid sequence although the Sce VMA endonuclease domain is reported to begin 27 amino acids N-terminal to intein block C (Duan *et al.*, 1997). However, based on comparison with the hedgehog processing domain and the analysis of Hall *et al.* (1997), PI-PspI residues 250–538 may not contain the entire endonuclease domain and do not contain the proposed DNA recognition region immediately following block B.

Splicing and cleavage with individual fragments or fragment pairs resulting in gaps

Previous studies indicated that cleavage at either Psp Pol-1 intein splice site does not require the conserved residues at the opposite splice junction (Xu and Perler, 1996). All six individual fragments were therefore tested for the ability to induce cleavage at the single splice junction present in that fragment. However, no cleavage was observed with any single fragment under any condition tested (Figure 7 and data not shown).

Several lines of evidence indicate that the splicing domain is limited to the terminal regions of the intein and that an endonuclease domain is inserted between the N- and C-terminal splicing subdomains (Pietrokovski, 1994, 1996; Chong and Xu, 1997; Derbyshire *et al.*, 1997; Duan *et al.*, 1997; Hall *et al.*, 1997; Perler *et al.*, 1997a; Telenti *et al.*, 1997). Therefore, fragment pairs resulting in gaps or deletions of intein sequence were tested for the ability to splice. All combinations of fragments that resulted in a gap in the intein (MI_{N1} plus $I_{C1}P$, $I_{C2}P$ or $I_{C3}P$ and MI_{N2} plus $I_{C3}P$) failed to splice or cleave (Figure 5, right panel and data not shown). To eliminate the possibility that the failure to splice was due to a failure to reassociate, MIP deletions were made to mimic these gaps, generating MIΔP precursors missing intein residues 109–440, 150–440, 251–440 or 273–440 and containing a 4–79 amino acid flexible linker at the deletion site. Splicing was performed under standard *cis*-splicing conditions (Xu *et al.*, 1993) or *trans*-splicing conditions. No splicing was observed with any precursor containing a deletion (Figure 5, right panel and data not shown), indicating that the failure of the gapped pairs to splice appears to be unrelated to splicing *in trans*.

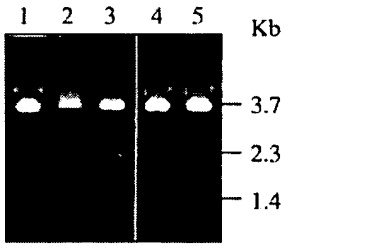


Fig. 6. Intein reconstitution also re-establishes PI-PspI endonuclease activity. Endonuclease activity of MIP fragments was assayed on pAKR7, which is a 3.7 kb plasmid containing a single PI-TliII (PI-PspI) site. Only *cis*-spliced intein from MIP52 and MI_{N2} plus $I_{C2}P$ reconstituted using protocol 1 were able to cleave the linearized plasmid into 2.3 and 1.4 kb pieces. Lane 1, MI_{N2} plus $I_{C2}P$ directly added to the endonuclease reaction without pre-treatment in urea; lane 2, MI_{N2} plus $I_{C2}P$ after protocol 1 treatment; lane 3, MIP52; lane 4, MI_{N2} after protocol 1 treatment; lane 5, $I_{C2}P$ after protocol 1 treatment.

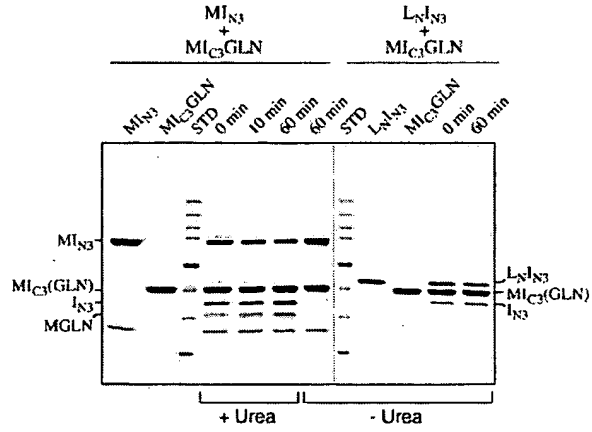


Fig. 7. *Trans*-splicing with other exteins. Splicing *in trans* of non-MIP precursors was examined after pre-incubation in 3.6 M urea buffer (protocol 2, +Urea lanes) or without pre-incubation in urea-containing buffers (protocol 4, –Urea lanes) prior to incubation in splicing buffer for 0–60 min at 37°C. Left panel: the C-extein (P) of $MI_{C3}P$ was replaced with the tripeptide, Gly–Leu–Asn (GLN), and $MI_{C3}GLN$ was reacted with MI_{N3}. The single substrate fragments were also incubated by themselves for 60 min (protocol 2). Splicing was assessed by quantifying the appearance of spliced products (MGLN, I_{N3} and MI_{C3}). Note that $MI_{C3}GLN$ and MI_{C3} co-migrate in this gel system and thus $MI_{C3}(GLN)$ represents either or both fragments. Some splicing products (I_{N3} and MGLN) are already observed after overnight pre-incubation in urea (0 min). Right panel: the N-extein (M) of MI_{N3} was replaced with the Lck fragment, L_N. L_NI_{N3} and $MI_{C3}GLN$ were pre-incubated in splicing buffer at 4°C for 30 min without pre-treatment in any urea-containing buffers and then incubated at 37°C for 60 min. The single fragments were also incubated by themselves at 37°C for 60 min. Since the L_NGLN spliced product is too small to be observed in this gel system and since L_N stains poorly with Coomassie blue, intein activity was assessed by the conversion of L_NI_{N3} to I_{N3}, which requires N-terminal cleavage.

Splicing with overlapping fragments or purification tags at the split sites

Previous studies indicated that, in a few cases, split proteins could contain one or more vector-derived amino acids or overlapping redundant sequences at the split site (Matsuyama *et al.*, 1990; Burbaum and Schimmel, 1991). Therefore, all possible pairs resulting in an overlap of intein sequence (MI_{N2} plus $I_{C1}P$, MI_{N3} plus $I_{C2}P$, MI_{N3} plus $I_{C1}P$) were assayed for the ability to splice. All overlapping pairs spliced, even though $I_{C1}P$ failed to splice with its complement, MI_{N1} (data not shown and Figure 5, left and center panels). In the case of MI_{N3} plus $I_{C1}P$, a functional intein was reconstituted despite the presence of 330 amino acids of redundant intein sequence. It is assumed that as the fragment pairs reassociated, redundant sequences were extruded from a location that did not interfere with formation of the splicing active site.

These data suggested that purification tags could be added to split sites. A six residue His tag was added to the split site of MI_{N3} and had no effect on splicing. MBP was fused to the split site of $I_{C3}P$ to yield $MI_{C3}P$ (Figure 2). The presence of an MBP (43 kDa) affinity tag in the I_C fragment, $MI_{C3}P$, had no effect on splicing efficiency (Figure 3 and Table I). However, the addition of MBP to the N-terminus of $I_{C3}P$ converted the insoluble $I_{C3}P$ fragment into a soluble $MI_{C3}P$ fragment. The presence of MBP at the N-terminus of fusion proteins often improves solubility of recombinant proteins in *E.coli*. Despite the fact that both MI_{N3} and $MI_{C3}P$ were synthesized as soluble proteins, no splicing was observed unless the fragments were pre-treated in 3.6 M urea (Figure 3).

Trans-splicing with other exteins

To test the general applicability of the *trans*-splicing system, other exteins were substituted for MBP and paramyosin in the MI_{N3} plus $MI_{C3}P$ system. In L_NI_{N3}, the N-extein (M) of MI_{N3} was replaced by a 9 kDa fragment of Lck tyrosine kinase encoding residues 52–121 (Perlmutter *et al.*, 1988). L_NI_{N3} plus either $MI_{C3}P$ or $I_{C3}P$ spliced after pre-incubation in 3.6 M urea (data not shown). In $I_{C3}L_C$, the paramyosin domain of $I_{C3}P$ was replaced by a 10 kDa fragment of Lck tyrosine kinase encoding residues 122–226 (Perlmutter *et al.*, 1988). Both MI_{N3} plus $I_{C3}L_C$ or L_NI_{N3} plus $I_{C3}L_C$ were able to splice if pre-treated in 3.6 M urea (data not shown). The paramyosin domain of $MI_{C3}P$ was then replaced by three amino acids (Gly–Leu–Asn or GLN), yielding $MI_{C3}GLN$. When $MI_{C3}GLN$ was mixed with a complementary fragment, MI_{N3}, splicing products MGLN, I_{N3} and MI_{C3} were observed only after pre-treatment in 3.6 M urea (Figure 7, left panel). Note that MI_{C3} and $MI_{C3}GLN$ are indistinguishable on these gels.

Two complementary fragments containing the smallest exteins, L_NI_{N3} and $MI_{C3}GLN$, were mixed directly in splicing buffer (Figure 7, right panel). Since the spliced exteins (LGLN) are too small to be clearly observed and since the Lck fragment stains poorly with Coomassie blue, reactions were scored positive for intein activity (splicing or cleavage) if L_NI_{N3} was converted to I_{N3} with time. As a control, L_NI_{N3} and $MI_{C3}GLN$ were also incubated as above in the absence of the complementary fragment. I_{N3} only accumulated when L_NI_{N3} was mixed with $MI_{C3}GLN$, indicating that the intein was active without pre-treatment in urea. However, the reaction under native conditions was not very efficient.

Discussion

Precursor fragments split within the Psp Pol-1 intein can be synthesized in separate hosts and then assembled *in vitro* to generate a fully active intein with both protein splicing and endonuclease activities. Remarkably, the intein fragments were able to assemble at temperatures of up to 100°C below their normal synthesis temperature. Once reconstituted, the active intein directed ligation of several test exteins including the chimera MP and L_N plus L_C, a Lck tyrosine kinase fragment spanning amino acids 52–226. *Trans*-splicing time courses confirmed the protein splicing pathway (Xu and Perler, 1996) since N-terminal and C-terminal cleavage could be monitored by the release of distinguishable intein fragments. Two of the three intein fragmentation sites yielded complementary precursor fragments that were capable of splicing, which is similar to previous data with other split proteins where only a fraction of the complementary pairs are able to reassemble (Matsuyama *et al.*, 1990; Burbaum and Schimmel, 1991). All combinations of overlapping fragments were also able to assemble into an active intein, including pairs with 330 amino acids of redundant intein sequence, requiring sufficient flexibility to displace these extra residues while forming the functional intein. The ability of these fragments to accommodate overlapping sequences suggested that they might be able to accommodate affinity tags at the split sites. The addition of a His tag to the C-terminus of I_N fragments had no effect on splicing, nor did the presence of a 43 kDa MBP affinity tag at the split site of $I_{C3}P$ have any effect on splicing, although it converted a previously insoluble fragment into a soluble fragment. No fragment pair resulting in a gap in the intein sequence and no MIP precursor containing a deletion mimicking these gaps was able to splice, suggesting that the failure of the gap pairs to splice is unrelated to the fragment assembly process.

Expression of MIP precursors is often accompanied by some degree of N-terminal or C-terminal cleavage *in vivo* (Xu *et al.*, 1993; Xu and Perler, 1996), but no *in vivo* or *in vitro* cleavage was observed with any individual fragment. This was unexpected since previous studies (Xu and Perler, 1996) had indicated that conserved Psp Pol-1 intein splice junction residues at one splice site are not required for cleavage at the opposite site and since even the smallest fragments contain all of the putative N-terminal (blocks A and B) or C-terminal (blocks F and G) splicing motifs (Pietrokovski, 1996; Duan *et al.*, 1997; Hall *et al.*, 1997; Perler *et al.*, 1997a; Telenti *et al.*, 1997). These results suggest that residues from both the N- and C-terminal subdomains of the Psp Pol-1 intein are required for cleavage at both splice junctions. This hypothesis is supported by the structures of the Sce VMA intein and hedgehog protein autoprocessing domain which indicate that the N- and C-terminal intein subdomains fold together to form one intermingled splicing domain.

This study also examined the effect of urea on the splicing reaction. Urea was required to solubilize the insoluble I_CP fragments. With all but the smallest exteins, pre-incubation in urea buffers was required for reconstitu-

tion of intein activity. Up to 6.0 M urea in the splicing buffer had no effect on splicing efficiency, but splicing was inhibited in 8.0 M urea. Splicing requires bringing at least three separate intein regions together to form the splicing active site (Ser1, His96, Asn537 and Ser538 in Psp Pol-1 intein) and splicing *in vitro* with MIP is very slow (half-time of 20–30 min, Xu and Perler, 1996). These results therefore suggest that urea concentrations of 6.0 M or less are not fully denaturing the intein since it is unlikely that a fully denatured intein could bring the above residues into proximity long enough for the slow *in vitro* splicing reaction to occur. These results suggest that the mechanisms resulting in thermostability of this protein from an extreme thermophile may also make it resistant to other forms of denaturation such as urea. In the future, it should be possible to maximize splicing of a split intein precursor containing any extein by determining optimum conditions for fragment solubility and precursor assembly. However, since the presence of the intein may prevent proper folding of the foreign extein protein, it may have to be denatured and renatured after splicing.

The cleavage and reassembly of protein splicing precursors opens up new avenues of protein analysis. Although the reconstituted precursor contains a break in the peptide backbone in the intein domain, after splicing the exteins are covalently linked with a native peptide bond (Cooper *et al.*, 1993). *Trans*-splicing of split precursors can thus be used to label or modify only a portion of the intact extein product. For example, an N-terminal fragment can be isolated from *E.coli* grown in media enriched with ^{13}C or ^{15}N. After splicing, the intact protein would only be labeled in the region of the N-extein. Such a partially labeled protein could be used to simplify structural determination by NMR analysis or possibly to allow the determination of larger protein structures. Another potential use of splicing with split inteins would be the modification by glycosylation, phosphorylation, dephosphorylation, etc. of only a subset of sites in a protein to determine which post-translationally modified site is important for enzyme activity. Finally, *trans*-splicing allows overexpression of highly toxic proteins, since no single cell contains the entire toxic protein.

Materials and methods

General procedures

Protein samples were electrophoresed on either 4–20% SDS–PAGE gels (Daiichi) or 12% SDS–PAGE gels (Novex). Protein concentrations were determined by the Bradford assay (Bio-Rad). Western blots were probed with anti-MBP, anti-*D.immitis* paramyosin or anti-Psp Pol-1 intein sera as previously described (Xu *et al.*, 1993). Western blot data are not shown, but were used to confirm the composition of all splicing and cleavage products. Protein Marker, Broad Range (New England Biolabs) standards were used. All cloning enzymes and oligonucleotides were from New England Biolabs and were used according to the manufacturer's instructions. All PCR fragments were sequenced in both directions by the New England Biolabs DNA sequencing core facility.

Construction of pMI_{N1}, pMI_{N2} and pMI_{N3}

DNA fragments encoding all or part of MBP and the indicated N-terminal fragments of the Psp Pol-1 intein were synthesized by PCR from pMIP21 (Figure 2) (Xu *et al.*, 1993). In each case, a stop codon and restriction enzyme site were introduced after the last intein codon. However, no extra residues were present at the split site, unless added later as an affinity tag. The fragment containing MBP and the first 249 codons of the intein was ligated into pAII17 (Hodges *et al.*, 1992) yielding pMI_{N2}, and fragments containing the first 108 or 440 intein codons were ligated into *Xho*I–*Bam*HI-digested pMI_{N2} yielding pMI_{N1} and pMI_{N3}, respectively. The primers were: 5′-GGAATTCCATATGAA-AATCGAAGAAGGT-3′ (pMI_{N2}); NEB 1237 (pMI_{N1} and pMI_{N3}), 5′-GGTCGTCAGACTGTCGATGAAGCC-3′, (pMI_{N1}), 5′-GGGGGATC-CTTACTCAACGAGATCCCCGTTCCTAT-3′; (pMI_{N2}), 5′-CGGGAT-CCCGTTATAGTGAGATAACGTCCCG-3′; and (pMI_{N3}), 5′-ATTGGATCCTTATCTGTATTCCGTAAACTTA-3′. PCR mixtures contained Vent DNA polymerase buffer (New England Biolabs), 0.2 mM each dNTP, 0.4 μM primers, 100 ng of plasmid DNA and 1 U of Vent DNA polymerase (New England Biolabs) in a 100 μl reaction. Amplification was carried out using a Perkin-Elmer Cetus 480 thermal cycler at 94°C for 30 s, 52°C for 30 s and 72°C for 135 s for 15–17 cycles. A C-terminal six amino acid His tag was added to the split site of MI_{N3} and MI_{N2} by standard procedures.

Construction of $pI_{C1}P$, $pI_{C2}P$, $pI_{C2}P\Delta$, $pI_{C3}P$ and $pMI_{C3}P$

Clones containing C-terminal fragments beginning at intein amino acids Thr110 ($pI_{C1}P$), Met250 ($pI_{C2}P$) or Lys441 ($pI_{C3}P$) were constructed by PCR from pMIP21 as described above (Figure 2). The $I_{C2}P$ fragment was subcloned into a pAII17 derivative, yielding a six residue His tag at the end of the paramyosin gene. The remaining two clones were generated by replacing the I_{C2} sequence in *Nde*I–*Bam*HI-digested $pI_{C2}P$ with PCR products. The primers used were 5′-GGGCATATGACTGGG-GAGGATGTCAAAATT-3′ ($pI_{C1}P$); 5′-GGAATTCCATATGCCAGA-GGAAGAACTG-3′ ($pI_{C2}P$); 5′-GAACATATGAAGAAAAAGAA-TGTATATCACTCTC-3′ ($pI_{C3}P$); 5′-ATAGTTTAGCGGCCGCTCAC-GACGTTGTAAAACG-3′ ($pI_{C2}P$); and 5′-GGGGGATCCAAAGCCA-GCAAGGAAATTCTC-3′ ($pI_{C1}P$ and $pI_{C3}P$). An 11 kDa (118 codon) C-terminal deletion was generated in $I_{C2}P\Delta$ to ease analysis since MI_{N2} and MP co-migrate on SDS–PAGE; this deletion had no effect on splicing. $pI_{C2}P$ was digested with *Nsi*I and *Sal*I, and blunted with T4 DNA polymerase prior to ligation. An *Nde*I–*Pst*I fragment from $pI_{C3}P$ (encoding $I_{C3}P$) was cloned into the *Eco*RI–*Pst*I site of pMAL-c2 (New England Biolabs) to form an in-frame fusion of MBP with $I_{C3}P$, yielding $pMI_{C3}P$.

Construction of L_NI_{N3}, $MI_{C3}L_C$ and $MI_{C3}GLN$

The L_NI_{N3} Lck tyrosine kinase fusion, encoding Lck residues 52–121, was produced by subcloning the PCR product amplified from the Lck cDNA clone, pCDNA1 (Perlmutter *et al.*, 1988), in place of MBP in pMI_{N3}His. PCR primers were 5′-GCTTACGCATATGGGCTCCAATC-CGCCGGCT-3′ and 5′-AGTGGTACCCATTCTTCCGGTAAAATGC-TGTTCGCTTTGGCCACAAA-3′. The *Nde*I–*Kpn*I-digested PCR products were gel purified and ligated with pMI_{N3} digested with the same enzymes.

The L_C fragment encoding Lck tyrosine kinase residues 122–226 was generated as above by PCR using primers 5′-GCGGATCCCTCTATGCA-CATAATAGCCTGGAGCCCGAACCC-3′ and 5′-GGGCGAAGCTTA-CTGGCAGGGGCGGC-3′. To replace paramyosin with L_C, the PCR product and $pMI_{C3}P$ were digested with *Bam*HI–*Hin*dIII and ligated as above. To replace paramyosin with the tripeptide Gly–Leu–Asn, $pMI_{C3}P$ was digested with *Bam*HI–*Hin*dIII, and a double-stranded oligonucleotide cassette (5′-GATCCCTCTATGCACATAATTCAGGCCTCAATTAA-3′ and 5′-AGCTTTAATTGAGGCCTGAATTATGTGCATAGAGG-3′) was ligated to the cut plasmid as above.

Expression and purification

pMI_{N1}, pMI_{N2}, pMI_{N3}, $pMI_{C3}P$, $pMI_{C3}L_C$ and $pMI_{C3}GLN$ were transformed into either ER2520 (*E.coli* B F-IDE3 (=I *sBam*HIo Δ*Eco*RI-B *int::lacI::Plac*UV5::T7 *gene1 imm21 Δnin5*) Δ*(mcrC-mrr)102::Tn10 gal ompT [lon, dcm]* [pLysS: CmR orip15A T7 gene 2]) or ER2538 (*E.coli* B; F- IDE3(=I *sBam*HIo Δ*Eco*RI-B *int::lacI*::*Plac*UV5::T7 *gene1 imm21 Δnin5*) *fhuA2 [lon] ompT gal sulA11 Δ(mcrC-mrr)114::IS10 R(mcr-73::miniTn10*–TetS*)2 endA1 R(zgb210::Tn10*–TetS*)* [pLysS: CmR orip15A T7 gene 2]) (Elisabeth Raleigh, New England Biolabs) and grown at 30°C in LB medium plus 100 μg/ml of ampicillin to an OD_{600} of ~0.5. The culture was induced with 0.4 mM isopropyl-β-D-thiogalactopyranoside (IPTG) and incubated overnight. The cells were sonicated in 50 ml of amylose column buffer (20 mM $NaPO_4$, pH 8.0, 0.5 M NaCl, 1.0 mM Na_2EDTA) and purified over an amylose column as described by the manufacturer (New England Biolabs).

$pI_{C2}P$, $pI_{C2}P\Delta$, $pI_{C1}P$, $pI_{C3}P$ and pL_NI_{N3} were transformed into ER2520 and induced as above, except $pI_{C1}P$ was induced with 0.04 mM IPTG for 2 h ($I_{C1}P$). Since a His tag was present after the paramyosin domain or the C-terminus of L_NI_{N3}, frozen cells were lysed by sonication in Ni^{2+} binding buffer (20 mM Tris–HCl pH 7.9, 500 mM NaCl, 16 mM

imidazole) and centrifuged at 10 000 *g* for 30 min. Insoluble I_CP proteins were solubilized in Ni^{2+} binding buffer containing 6.0 M urea and purified over a Ni^{2+} charged column in the presence of 6.0 M urea (Novagen, 5 ml of resin) as described by the manufacturer. Alternatively, insoluble proteins were purified by merely washing the pellet several times with Ni^{2+} binding buffer containing 1.0 M urea prior to resuspension in 6.0 M urea.

Fragment assembly (trans-splicing) protocols and quantitation

Coomassie blue-stained gels were digitized with a Microtek Scanmaker III and analyzed with NIH Image 1.51 software for quantitation. Samples were not normalized for protein loss during the time course since there appeared to be unequal loss of each substrate and product. However, it should be noted that there was up to 35% loss of total protein during most splicing reactions of 1.5 h or more, indicating precipitation during the reaction, especially of I_C-containing fragments which are known to precipitate with time in the absence of urea. There may also be differences in staining of fragments with Coomassie blue since splicing of MIP *in cis* yields only 71% as many moles of MP product as I product, with no compensating increase in M or P cleavage products. Therefore, splicing efficiency was calculated to reflect actual yields of spliced product (the product of interest) by dividing the moles of spliced product at the end of the reaction by the initial moles of limiting substrate (usually MI_N) and multiplying by 100.

Purified protein fragments were stored in either amylose elution buffer, Ni^{2+} elution buffer (± 6.0 M urea) or exchanged into buffer E [50 mM Tris–HCl pH 7.5, 5% acetate 0.1 mM EDTA, 1 mM dithiothreitol (DTT), 140 mM β-mercaptoethanol and 7.2 M urea, equilibrated to pH 7.5]. The proteins were then combined to a final concentration of 0.5–2.5 mg/ml. Several pre-incubation protocols were examined involving treatment alone or after mixing in various concentrations of urea (0–8.0 M) and different buffers for differing times at 4 or 37°C prior to dilution (1:10 to 1:50) into splicing buffer (20 mM $NaPO_4$, pH 6, 0.5 M NaCl, 1 mM EDTA). However, most experiments presented in this study were performed with one of the following standard protocols (Table I). In protocol 1, both fragments were in buffer E (7.2 M urea) or Ni^{2+} elution buffer (6.0 M urea) prior to mixing. In protocol 2, MI_N in amylose elution buffer was combined with an equal volume of C-terminal fragment in buffer E or Ni^{2+} elution buffer, resulting in a 3.6 or 3.0 M urea pre-incubation buffer, respectively. In protocol 3, fragments were mixed in various concentrations of urea (0–8.0 M) and then diluted into splicing buffer containing the same concentration of urea. In protocol 4, the fragments were mixed in amylose elution buffer without urea and diluted into splicing buffer without urea. In protocol 5, MI_N in amylose elution buffer was combined with I_CP in buffer E (7.2 M urea) and varying amounts of amylose elution buffer to yield pre-incubation buffers containing 0–3.6 M urea and then diluted 10-fold into splicing buffer without urea. All fragment mixtures except those using protocol 4 were pre-incubated in urea-containing buffers for 0 h to overnight at 4°C. Following this pre-incubation step, the mixtures were diluted 10- to 50-fold in splicing buffer and immediately incubated at 37°C for 0–24 h to stimulate the splicing reaction. Most experiments presented were diluted 10-fold since no significant difference was observed with higher dilutions. Purified $I_{C1}P$, $I_{C2}P$, $I_{C3}P$ and $I_{C3}L_C$ precipitated within 2–4 h after rapid dilution out of urea-containing buffers. Therefore, in most cases, splicing reactions were limited to 2 h to avoid loss of these substrates due to precipitation rather than splicing.

In some experiments, the following folding aids were added to either or both the pre-incubation or splicing buffers (protocol 2): Triton X-100 (1%), glycerol (10%), PEG 8000 (3 mg/ml), arginine (0.5 M) and SDS (1 mg/ml and 0.1 mg/ml).

Endonuclease assays

pAKR7 contains the 714 bp *Eco*RI fragment from pAKK4 (Hodges *et al.*, 1992) in Bluescript SK– and encodes a single PI-TliII recognition site created by the deletion of the *Tli pol-1 intein* from the *Thermococcus litoralis* DNA polymerase gene. This site is similar, but not identical, to the PI-PspI recognition site predicted by deletion of the *Psp pol-1 intein*. Although PI-TliII and PI-PspI are isoschizomers (F.Perler, unpublished data), it is not known whether they cut each other's recognition site with the same efficiency. Digestions were performed for 1 h at 50°C in 0.1 M NaCl, 50 mM Tris–HCl, 10 mM $MgCl_2$, 1 mM DTT, pH 8.6 (at 25°C). Similar amounts of PI-PspI were present in each reaction in the form of (i) MIP52 (>80% present as the MIP precursor at the beginning of the reaction) which contains an insert of MILVA prior to Ser1 of MIP, (ii) single fragments MI_{N2} or $I_{C2}P$, (iii) MI_{N2} and $I_{C2}P$ fragment pairs that had been pre-assembled using protocol 1 with a 2 h splicing reaction or (iv) MI_{N2} and $I_{C2}P$ directly added to the digestion mixture. The 3.7 kb pAKR7 DNA was linearized with *Xmn*I prior to digestion with PI-PspI so that digestion with PI-PspI would yield fragments of 1.4 and 2.3 kb.

Construction of intein deletions

The insert in pMI_{N1} was amplified by PCR from the *Nde*I site at the beginning of the *malE* gene to the end of I_{N1} with the addition of an *Spe*I and *Nde*I site at the 3′ end. This fragment was digested with *Nde*I and subcloned into the *Nde*I site of $pI_{C3}P$ to create pMIΔ109–440P which contains a *Spe*I and *Nde*I site between the intein sequences resulting in the insertion of four amino acids (Thr–Ser–His–Met). Two complementary primers (5′-CT AGG GGC TAT GAC CTG CCC ATG GTT GAG GAA GGA GAG CCT GAC C-3′ and 3′-C CCG ATA CTG GAC GGG TAC CAA CTC CTT CCT CTC GGA CTG GGA TC-5′) were ligated into *Spe*I-digested pMIΔ109–440P, resulting in the insertion of 15 amino acids at the deletion site per copy of double-stranded linker. The linker can be inserted in either direction, resulting in different sequence combinations. All subsequent MIP deletion clones were made by digesting pMIΔ109–440P with *Eco*RI–*Spe*I and substituting similarly digested PCR products. MIP deletion precursors were purified as described above for MI_N proteins. MIΔ109–440P precursors containing 1–5 copies of the linker, MIΔ150–440P and MIΔ273–440P precursors containing 1–3 copies of the linker, and MIΔ251–440P precursors containing one or two linkers were all tested for the ability to splice under standard *cis*- (Xu and Perler, 1996) and *trans*-splicing conditions.

Acknowledgements

We gratefully acknowledge Jonathan Lee of Boston University for support of R.A.K. and for helpful discussions. We thank Bill Jack for the gift of pAKR7 DNA and Jack Benner and Ted Davis for sharing the results of their Psp Pol-1 intein protease digestion studies. We thank Chris Noren for helpful discussions and Don Comb for support and encouragement. Sections of this study were performed at New England Biolabs as part of the thesis requirements of E.A. and D.P. at the Ecole Superieure de Biotechnologie de Strasbourg, France (1994–1995).

References

Anfinsen,C.B. and Scheraga,H.A. (1975) Experimental and theoretical aspects of protein folding. *Adv. Protein Chem.*, **29**, 205–300.

Bremer,M., Gimble,F.S., Thorner,J. and Smith,C. (1992) VDE endonuclease cleaves *Saccharomyces cerevisiae* genomic DNA at a single site: physical mapping of the VMA1 gene. *Nucleic Acids Res.*, **20**, 5484.

Burbaum,J.J. and Schimmel,P. (1991) Assembly of a class I tRNA synthetase from products of an artificially split gene. *Biochemistry*, **30**, 319–324.

Chong,S. and Xu,M.Q. (1997) Protein splicing of the *Saccharomyces cerevisiae* VMA intein without the endonuclease motifs. *J. Biol. Chem.*, **272**, 15587–15590.

Chong,S., Shao,Y., Paulus,H., Benner,J., Perler,F.B. and Xu,M.Q. (1996) Protein splicing involving the *Saccharomyces cerevisiae* VMA intein. The steps in the splicing pathway, side reactions leading to protein cleavage and establishment of an *in vitro* splicing system. *J. Biol. Chem.*, **271**, 22159–22168.

Cooper,A.A., Chen,Y., Lindorfer,M.A. and Stevens,T.H. (1993) Protein splicing of the yeast *TFP1* intervening protein sequence: a model for self-excision. *EMBO J.*, **12**, 2575–2583.

Derbyshire,V., Wood,D.W., Wu,W., Dansereau,J.T., Dalgaard,J.Z. and Belfort,M. (1997) Genetic definition of a protein-splicing domain: functional mini-inteins support structure predictions and a model for intein evolution. *Proc. Natl Acad. Sci. USA*, **94**, 11466–11471.

Duan,X., Gimble,F.S. and Quiocho,F.A. (1997) Crystal structure of PI-SceI, a homing endonuclease with protein splicing activity. *Cell*, **89**, 555–564.

Gross,M., Wyss,M., Furter-Graves,E.M., Wallimann,T. and Furter,R. (1996) Reconstitution of active octameric mitochondrial creatine kinase from two genetically engineered fragments. *Protein Sci.*, **5**, 320–330.

Hall,T.M.T., Porter,J.A., Young,K.E., Koonin,E.V., Beachy,P.A. and Leahy,D.J. (1997) Crystal structure of a hedgehog autoprocessing domain: conservation of structure, sequence and cleavage mechanism between hedgehog and self-splicing proteins. *Cell*, **91**, 85–97.

Hodges,R.A., Perler,F.B., Noren,C.J. and Jack,W.E. (1992) Protein splicing removes intervening sequences in an archaea DNA polymerase. *Nucleic Acids Res.*, **20**, 6153–6157.

Kanaya,E. and Kanaya,S. (1995) Reconstitution of *Escherichia coli* RNase HI from the N-fragment with high helicity and the C-fragment with a disordered structure. *J. Biol. Chem.*, **270**, 19853–19860.

Kato,I. and Anfinsen,C.B. (1969) On the stabilization of ribonuclease S-protein by ribonuclease S- peptide. *J. Biol. Chem.*, **244**, 1004–1007.

Kawasaki,M., Nogami,S., Satow,Y., Ohya,Y. and Anraku,Y. (1997) Identification of three core regions essential for protein splicing of the yeast Vma1 protozyme. *J. Biol. Chem.*, **272**, 15668–15674.

Matsuyama,S., Kimura,E. and Mizushima,S. (1990) Complementation of two overlapping fragments of SecA, a protein translocation ATPase of *Escherichia coli*, allows ATP binding to its amino-terminal region. *J. Biol. Chem.*, **265**, 8760–8765.

Mueller,J.E., Bryk,M., Loizos,N. and Belfort,M. (1994) Homing endonucleases. In Linn,S.M. Lloyd,R.S. and Roberts,R.J. (eds), *Nucleases*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, pp. 111–143.

Perler,F.B. *et al.* (1992) Intervening sequences in an Archaea DNA polymerase gene. *Proc. Natl Acad. Sci. USA*, **89**, 5577–5581.

Perler,F.B., Davis,E.O., Dean,G.E., Gimble,F.S., Jack,W.E., Neff,N., Noren,C.J., Thorner,J. and Belfort,M. (1994) Protein splicing elements: inteins and exteins—a definition of terms and recommended nomenclature. *Nucleic Acids Res.*, **22**, 1125–1127.

Perler,F.B., Olsen,G.J. and Adam,E. (1997a) Compilation and analysis of intein sequences. *Nucleic Acids Res.*, **25**, 1087–1093.

Perler,F.B., Xu,M.-Q. and Paulus,H. (1997b) Protein splicing and autoproteolysis mechanisms. *Curr. Opin. Chem. Biol.*, **1**, 292–299.

Perlmutter,R.M., Marth,J.D., Lewis,D.B., Peet,R., Ziegler,S.F. and Wilson,C.B. (1988) Structure and expression of lck transcripts in human lymphoid cells. *J. Cell. Biochem.*, **38**, 117–126.

Pietrokovski,S. (1994) Conserved sequence features of inteins (protein introns) and their use in identifying new inteins and related proteins. *Protein Sci.*, **3**, 2340–2350.

Pietrokovski,S. (1996) A new intein in Cyanobacteria and its significance for the spread of inteins. *Trends Genet.*, **12**, 287–288.

Rost,B., Sander,C. and Schneider,R. (1994) PHD—an automatic mail server for protein secondary structure prediction. *CABIOS*, **10**, 53–60.

Sancho,J. and Fersht,A.R. (1992) Dissection of an enzyme by protein engineering. The N and C-terminal fragments of barnase form a native-like complex with restored enzymatic activity. *J. Mol. Biol.*, **224**, 741–747.

Shao,Y., Xu,M.Q. and Paulus,H. (1995) Protein splicing: characterization of the aminosuccinimide residue at the carboxyl terminus of the excised intervening sequence. *Biochemistry*, **34**, 10844–10850.

Shao,Y., Xu,M.-Q. and Paulus,H. (1996) Protein splicing: evidence for an *N–O* acyl rearrangement as the initial step in the splicing process. *Biochemistry*, **35**, 3810–3815.

Tasayco,M.L. and Chao,K. (1995) NMR study of the reconstitution of the beta-sheet of thioredoxin by fragment complementation. *Proteins*, **22**, 41–44.

Telenti,A., Southworth,M., Alcaide,F., Daugelat,S., Jacobs,W.R. and Perler,F.B. (1997) The *Mycobacterium xenopi* GyrA protein splicing element: characterization of a minimal intein. *J. Bacteriol.*, **179**, 6378–6382.

Xu,M. and Perler,F.B. (1996) The mechanism of protein splicing and its modulation by mutation. *EMBO J.*, **15**, 5146–5153.

Xu,M., Southworth,M.W., Mersha,F.B., Hornstra,L.J. and Perler,F.B. (1993) *In vitro* protein splicing of purified precursor and the identification of a branched intermediate. *Cell*, **75**, 1371–1377.

Xu,M., Comb,D.G., Paulus,H., Noren,C.J., Shao,Y. and Perler,F.B. (1994) Protein splicing: an analysis of the branched intermediate and its resolution by succinimide formation. *EMBO J.*, **13**, 5517–5522.

Received August 18, 1997; revised November 27, 1997; accepted December 17, 1997